EPFL

Thèse n° 9681

On matching data and model in LF-MMI-based dysarthric speech recognition

Présentée le 29 juin 2023

Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Enno HERMANN

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Dr J.-M. Odobez, Dr M. Magimai Doss, directeurs de thèse
Prof. E. Nöth, rapporteur
Prof. I. Trancoso, rapporteuse
Prof. J.-Ph. Thiran, rapporteur

École polytechnique fédérale de Lausanne

2023

Acknowledgements

First of all I am very grateful to my supervisor Dr. Mathew Magimai.-Doss for his invaluable advice and guidance throughout my PhD. He never failed to suggest new research directions, put my work into a larger context and keep me on the right track. I also thank the members of my thesis committee, Prof. Pascal Frossard, Dr. Jean-Marc Odobez, Prof. Elmar Nöth, Prof. Isabel Trancoso, and Prof. Jean-Philippe Thiran for their constructive feedback and comments.

My work was funded by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287 (TAPAS) and by the Innosuisse grant agreement No PFFS-21-47 (IICT flagship). I especially enjoyed the many opportunities the TAPAS network provided to travel and meet fellow researchers and cherished our project meetings.

I am grateful to Bence Halpern and Karl El Hajal for our collaboration on voice conversion for dysarthric speech, which served as the basis for some of my data augmentation experiments in Chapter 6.

The Covid-19 pandemic disrupted many plans, so my internship at Therapy Box could only be done remotely. Nonetheless, I learned a lot about the challenges of turning research into products and I thank Swapnil and Rebecca for this opportunity.

Thanks to all colleagues and friends at Idiap who really enriched my PhD experience in the office, but also outdoors while hiking or skiing. I would like to thank my family for always believing in me. Finally, thank you Verena for your constant love and support over all these years.

This document is based on the EPFL PolyDoc thesis template[1]. The CV is adapted from a template by LianTze Lim[2].

Lausanne, February 2023 Enno

[1] https://github.com/glederrey/EPFL_thesis_template

[2] https://www.overleaf.com/latex/templates/a-customised-curve-cv/mvmbhkwsnmwv

Abstract

In light of steady progress in machine learning, automatic speech recognition (ASR) is entering more and more areas of our daily life, but people with dysarthria and other speech pathologies are left behind. Their voices are underrepresented in the training data and so different from typical speech that ASR systems fail to recognise them. This thesis aims to adapt both acoustic models and training data of ASR systems in order to better handle dysarthric speech.
We first build state-of-the-art acoustic models based on sequence-discriminative lattice-free maximum mutual information (LF-MMI) training that serve as baselines for the following experiments. We propose the dynamic combination of models trained on either control, dysarthric, or both groups of speakers to address the acoustic variability of dysarthric speech. Furthermore, we combine models trained with either phoneme or grapheme acoustic units in order to implicitly handle pronunciation variants.
Second, we develop a framework to analyse the acoustic space of ASR training data and its discriminability. We observe that these discriminability measures are strongly linked to subjective intelligibility ratings of dysarthric speakers and ASR performance.
Finally, we compare a range of data augmentation methods, including voice conversion and speech synthesis, for creating artificial dysarthric training data for ASR systems. With our analysis framework, we find that these methods reproduce characteristics of natural dysarthric speech.

Keywords: automatic speech recognition, dysarthria, pathological speech processing, LF-MMI, acoustic subword units, data augmentation

Résumé

Grâce aux progrès constants de l'apprentissage automatique, la reconnaissance automatique de la parole pénètre de plus en plus dans notre vie quotidienne, mais les personnes atteintes de dysarthrie et d'autres pathologies de la parole sont laissées pour compte. Leurs voix sont sous-représentées dans les données d'entraînement et si différentes de la parole typique que les systèmes de reconnaissance vocale ne les reconnaissent pas. Cette thèse vise à adapter à la fois les modèles acoustiques et les données d'entraînement des systèmes de reconnaissance vocale afin de mieux traiter la parole dysarthrique.

Nous construisons d'abord des modèles acoustiques de pointe basés sur l'entraînement discriminatif des séquences LF-MMI, qui servent de base aux expériences suivantes. Nous proposons la combinaison dynamique de modèles entraînés soit sur des locuteurs typiques, soit sur des locuteurs dysarthriques, soit sur les deux groupes de locuteurs, afin d'aborder la variabilité acoustique de la parole dysarthrique. De plus, nous combinons des modèles entraînés avec des unités acoustiques de phonèmes ou de graphèmes afin de traiter implicitement les variantes de prononciation.

Deuxièmement, nous développons un cadre pour analyser l'espace acoustique des données d'entraînement et sa discriminabilité. Nous observons que ces mesures de discriminabilité sont fortement liées aux évaluations subjectives de l'intelligibilité des locuteurs dysarthriques et aux performances de reconnaissance vocale.

Enfin, nous comparons une série de méthodes d'augmentation de données, y compris la conversion de la voix et la synthèse vocale, pour créer des données d'entraînement dysarthriques artificielles pour les systèmes de reconnaissance vocale. Grâce à notre cadre d'analyse, nous constatons que ces méthodes reproduisent les caractéristiques de la parole dysarthrique naturelle.

Keywords : reconnaissance automatique de la parole, dysarthrie, traitement de la parole pathologique, LF-MMI, unités de sous-mots acoustiques, augmentation de données

Zusammenfassung

Angesichts des stetigen Fortschritts im maschinellen Lernen, hält automatische Spracherkennung Einzug in immer mehr Bereiche unseres täglichen Lebens, aber Personen mit Dysarthrie und anderen Sprechstörungen werden zurückgelassen. Ihre Stimmen sind nicht genügend in den Trainingsdatensätzen repräsentiert und so unterschiedlich, dass typische Spracherkennungssysteme sie nicht verstehen. Diese Dissertation zielt darauf ab, sowohl akustische Modelle, als auch die Trainingsdaten von Spracherkennungssystemen so anzupassen, dass sie dysarthrische Sprache besser verarbeiten können.
Zunächst entwickeln wir akustische Modelle basierend auf diskriminativem LF-MMI Training, die als Grundlinien für die folgenden Experimente dienen. Wir schlagen die dynamische Kombination von Modellen vor, die entweder mit typischen, dysarthrischen, oder beiden Sprachdaten trainiert worden sind, um die akustische Variabilität der dysarthrischen Sprache zu adressieren. Weiterhin kombinieren wir Modelle, die entweder mit phonemischen oder mit graphemischen Subworteinheiten trainiert worden sind, um implizit mit Aussprachevariationen umzugehen.
Zweitens entwickeln wir ein Rahmenwerk, um die akustischen Einheiten der Trainingsdaten für Spracherkennung und ihre Unterscheidbarkeit zu analysieren. Wir beobachten, dass diese Unterscheidbarkeitskriterien eng mit subjektiven Verständlichkeitsbewertungen der dysarthrischen Sprecher und den Ergebnissen der Spracherkennung zusammenhängen.
Abschliessend vergleichen wir unterschiedliche Datenaugmentationsmethoden, einschliesslich Sprachumwandlung und Sprachsynthese, um künstliche dysarthrische Trainingsdaten für Spracherkennungssysteme zu generieren. Mit unserem Analysesystem ermitteln wir, dass diese Methoden Eigenschaften natürlicher dysarthrischer Sprache reproduzieren.

Schlüsselwörter: automatische Spracherkennung, Dysarthrie, Verarbeitung pathologischer Sprache, LF-MMI, akustische Subwort-Einheiten, Datenaugmentation

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Dysarthria is a motor speech disorder caused by damage to the nervous system that results in a reduction of control over the muscles involved in speech production. This leads to atypical breathing, imprecise articulation, lower speaking rates, dysfluencies, and overall reduced speech intelligibility (Duffy, 2012). Common causes for dysarthria are stroke, cerebral palsy or neurodegenerative diseases, such as Parkinson's disease and amyotrophic lateral sclerosis (ALS).

Often these patients face difficulties not only with speech production, because other parts of their motor system are affected as well, creating challenges in carrying out everyday tasks. Assistive technology that recognises such pathological speech and is integrated with a home automation system could therefore help with daily tasks, such as switching on the light or changing TV channels, that are otherwise very difficult for people with limited motor control because input methods like buttons or touch screens are not designed for their needs.

Although considerable progress has been made in the field of automatic speech recognition (ASR), it has been found that current commercial and open-source ASR systems still perform poorly on atypical speech, including accented (Hinsvark et al., 2021), children's (Dubagunta et al., 2019) and pathological speech (Moore et al., 2018). On the other hand, humans struggle to understand severely dysarthric speech as well (Mengistu and Rudzicz, 2011b) and better ASR systems could serve as a communication aid for people not familiar with dysarthric speech.

This highlights the need for further research on dysarthric ASR that results in tangible improvements in mainstream speech technology and thus directly improves the quality of life for people with speech disorders.

1.2 Summary of contributions

The main contributions of this thesis are:

- **Sequence-discriminative ASR baselines**
  We build state-of-the-art acoustic models with sequence-discriminative lattice-free maximum mutual information (LF-MMI) training on the Torgo and UA-Speech corpora of dysarthric speech as baselines for the following experiments. We also propose a new evaluation protocol for the Torgo corpus that treats the isolated word and sentence recognition tasks separately.

- **Handling acoustic and lexical variability through model combination**
  We find that dynamic combination of acoustic models trained on different groups of speakers improves the handling of acoustic and lexical variability of dysarthric speech. Furthermore, the combination of models trained with different acoustic units, such as phonemes and graphemes, implicitly supports pronunciation variations and improves ASR results on dysarthric and children's speech.

- **Analysis framework to measure acoustic discriminability of speech**
  We develop an analysis framework to measure the acoustic discriminability of ASR training data based on Kullback-Leibler (KL) divergences between Gaussian distributions estimated for acoustic subword units. These measures are well correlated with subjective intelligibility ratings of dysarthric speech and the performance of dysarthric speech recognition systems.

- **Comparison of data augmentation approaches for dysarthric ASR**
  We conduct a detailed comparison of different data augmentation methods for dysarthric ASR, including voice conversion (VC) and text-to-speech (TTS). We evaluate the synthetic speech output from these systems with our proposed analysis framework to determine how well they reproduce characteristics of the original dysarthric speech.

1.3 Thesis outline

The remainder of this thesis is structured as follows:

Chapter 2 briefly summarises the relevant background on ASR and discusses previous approaches to dysarthric speech recognition. We also present the datasets used for the experiments in this thesis.

Sequence-discriminative acoustic models trained with the LF-MMI objective function are a state-of-the-art approach for ASR. In Chapter 3 we investigate why they are also very effective for dysarthric speech recognition and develop baseline models for the Torgo and UA-Speech corpora that will be used throughout this thesis.

Chapter 4 discusses how the combination of acoustic models trained on different groups of speakers or with different acoustic unit sets can handle the acoustic and lexical variability of dysarthric speech. We also evaluate this approach on children's and non-native children's speech as other forms of atypical speech.

In Chapter 5, we propose a framework for analysing and predicting the performance of dysarthric speech recognition systems based on the acoustic discriminability of the training data. We also relate this to the variations in intelligibility between different dysarthric speakers and to intelligibility ratings of TTS systems.

We then compare different data augmentation approaches for improving dysarthric speech recognition in Chapter 6 with this analysis framework: voice conversion based on signal processing and generative adversarial networks (GANs), and speech synthesis. We further investigate whether synthesised dysarthric speech could also be used in a few-shot setting to build ASR systems for new speakers from very little data.

Finally, Chapter 7 summarises this thesis and suggests directions for future work.

1.4 Peer-reviewed publications based on this thesis work

Chapter 3:

Hermann, E. and Magimai.-Doss, M. (2020). Dysarthric Speech Recognition with Lattice-Free MMI. In *Proceedings of ICASSP*, pages 6109–6113

Chapters 4 and 5:

Hermann, E. and Magimai.-Doss, M. (2021). Handling acoustic variation in dysarthric speech recognition systems through model combination. In *Proceedings of Interspeech*, pages 4788–4792

Chapter 6:

Halpern, B. M., Fritsch, J., Hermann, E., van Son, R., Scharenborg, O., and Magimai.-Doss, M. (2021). An Objective Evaluation Framework for Pathological Speech Synthesis. In *Proceedings of the ITG Conference on Speech Communication*

Hermann, E. and Magimai.-Doss, M. (2023). Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation. In *Proceedings of Interspeech (accepted)*

2 Background

This chapter introduces the relevant technical background for this thesis. We give a general overview of ASR in Section 2.1, summarise previous works on ASR for dysarthric speech in Section 2.2, and describe the datasets used throughout this thesis in Section 2.3.

2.1 Automatic speech recognition

ASR is the task of finding the most likely sequence of words $\boldsymbol{w}^* = (w_1, \ldots, w_M)$ that corresponds to a given speech signal, represented by a sequence of acoustic feature vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$:

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{X}) \,. \tag{2.1}$$

Following Bayes' theorem we can decompose this into an *acoustic model* $P(\boldsymbol{X}|\boldsymbol{w})$ and a *language model* $P(\boldsymbol{w})$ that can be trained separately:

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \frac{P(\boldsymbol{X}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{X})} \tag{2.2}$$

$$= \arg\max_{\boldsymbol{w}} P(\boldsymbol{X}|\boldsymbol{w})P(\boldsymbol{w}) \,. \tag{2.3}$$

The ASR output can be further processed, for example, a spoken language understanding system may translate the ASR output *"louder"* into an intent that leads to increasing the volume of a loudspeaker.

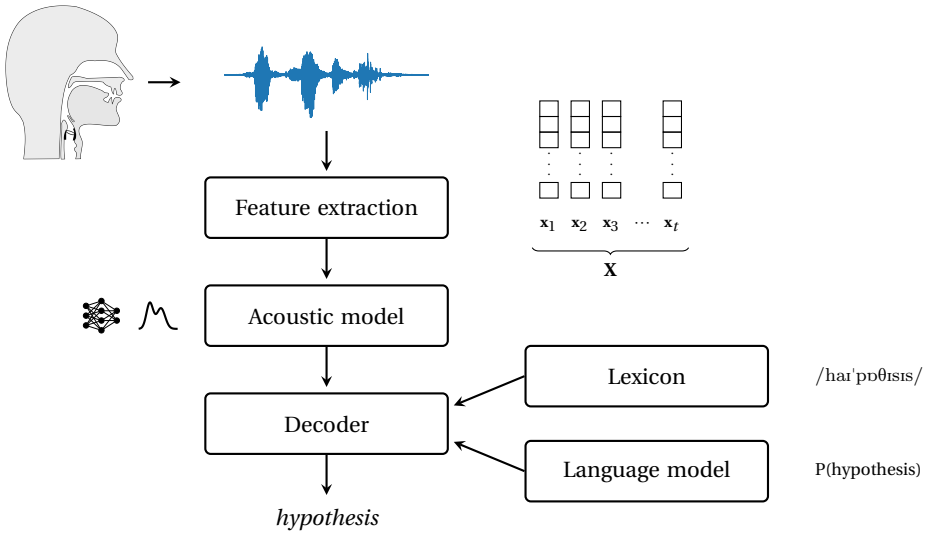


Figure 2.1: Overview of an ASR system to convert a speech signal $\boldsymbol{X}$ into a word hypothesis.

2.1.1 Overview

Figure 2.1 illustrates the general pipeline of an ASR system. First, the speech signal is converted into a sequence of acoustic feature vectors with a lower sampling rate. Traditionally, these are often Mel-frequency cepstral coefficients (MFCCs), but with the advent of neural networks, representations with fewer processing steps that could lead to information loss are increasingly preferred, such as Mel spectrograms or just windows of the raw signal itself (Palaz et al., 2019).

The acoustic model $P(\boldsymbol{X}|\boldsymbol{w})$, typically a Gaussian mixture model (GMM) or deep neural network (DNN), is trained on a database of transcribed speech recordings to distinguish between acoustic units (or states). The most common units are phonemes, context-dependent triphones or graphemes, exemplified in Table 2.1. Separately modelling triphones for each possible left and right phoneme context would lead to too many units, many of which would be seen only rarely or not at all in the data. In practice, acoustically similar context-dependent triphones are therefore clustered with a decision tree into *senones* (Young et al., 1994). Unless graphemes are used directly, a pronunciation lexicon, in conjunction with a grapheme-to-phoneme conversion system for out-of-vocabulary words, is additionally required to map from graphemes to phonemes.

Table 2.1: Different choices of acoustic units for the word *act*. For context-dependent triphones, the left phoneme context is prefixed with "-" and the right context suffixed with "+".

Phonemes	**Context-dependent triphones**	**Graphemes**
/æ/, /k/, /t/	/æ+k/, /æ-k+t/, /k-t/	a, c, t

A language model $P(\boldsymbol{w})$ is trained on text data to learn which words are likely to follow each other. No corresponding speech recordings are required for this text, so the language model

training data is usually much larger than what the acoustic model is trained on. For simple applications, it is also possible to manually create a grammar that describes the possible output sequences.

Finally, the *decoder* combines the outputs of the acoustic and language models and the lexicon and searches for the most likely transcription for the given input speech signal. The decoder is often based on weighted finite state transducers (Mohri et al., 2002) because they allow efficient combination of context-dependent hidden Markov model (HMM) models, the lexicon, and the language model into a single decoding graph.

In this thesis, we focus on improvements to the acoustic models and will not make any changes to the language models because we are working with small datasets that only contain a limited number of words. This also corresponds to the real-life use case of an assistive home automation system developed for a fixed set of commands. We refer the reader to Yue et al. (2020) for language modelling for unconstrained continuous dysarthric ASR and to Hernandez et al. (2022) for end-to-end training without any external language model.

The performance of ASR systems is generally evaluated with the word error rate (WER). For this, the number of insertions, deletions, and substitutions in the ASR output compared to the correct transcript are obtained with the *minimum edit distance* algorithm. The WER is then computed as follows:

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Number of words in the correct transcript}} \tag{2.4}$$

It is reported in percent, but values may be greater than one hundred if the number of insertions is high.

2.1.2 Acoustic model training

In this section, we describe common acoustic model training methods that we employ in this thesis. The acoustic units in ASR are typically modelled with HMMs whose state emission probabilities are estimated with GMMs or DNNs as illustrated in Figure 2.2.

The basic objective function for generative HMM/GMM acoustic model training is maximum likelihood (Rabiner, 1989; Hadian et al., 2018b)

$$\mathcal{F}_{ML} = \sum_{u=1}^{U} \log p_\theta(\boldsymbol{X}^u | \boldsymbol{Q}_{\boldsymbol{w}^u}) \tag{2.5}$$

$$= \sum_{u=1}^{U} \log \sum_{\boldsymbol{q} \in \boldsymbol{Q}_{\boldsymbol{w}^u}} \prod_{t=1}^{T_u} p(q_t | q_{t-1}) p(\boldsymbol{x}_t^u | q_t) \,, \tag{2.6}$$

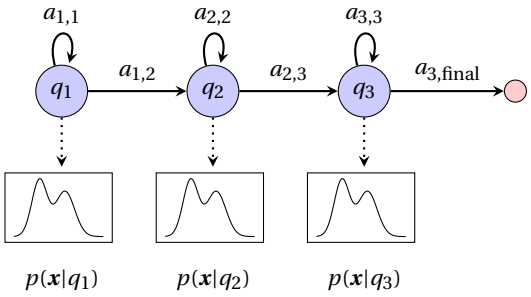


Figure 2.2: 3-state left-to-right HMM in an ASR system with transition probabilities $a_{i,j}$ from state q_i to q_j and emission probabilities $p(\boldsymbol{x}|q)$ modelled by GMMs.

where θ are the trainable HMM parameters, namely transition and emission probabilities, U is the number of utterances in the training data and $\boldsymbol{X}^u$ are the acoustic features of the u^{th} utterance of length T_u with corresponding transcription $\boldsymbol{w}^u$. $\boldsymbol{Q}_{\boldsymbol{w}^u}$ denotes all possible state sequences for transcription $\boldsymbol{w}^u$ – words with multiple possible pronunciations result in multiple state sequences.

Hybrid HMM/DNN models are trained discriminatively at the frame level, i.e. the DNN is trained to distinguish between acoustic units for a given frame of speech. This is achieved by maximising the negative cross-entropy (CE) (Bourlard and Morgan, 1994; Hadian et al., 2018b)

$$\mathcal{F}_{CE} = \sum_{u=1}^{U} \sum_{t=1}^{T_u} \log(y_t^u \cdot z_t^u) \,, \tag{2.7}$$

where y_t^u is the output of the neural network at time t, a probability distribution over HMM states, and z_t^u is a one-hot vector identifying the correct state at time t. This state alignment $\boldsymbol{z}^u$ is obtained from a previously trained HMM/GMM model.

On the other hand, lattice-free maximum mutual information (LF-MMI), and other MMI-based objective functions, are trained discriminatively at the sequence level. Originally developed for HMM/GMM model training (Bahl et al., 1986), they can be equally applied in the context of hybrid HMM/DNN ASR. In addition to maximising the likelihood of the correct state sequences, MMI minimises the likelihood of any other state sequence. This better matches the goal of finding the best output sequence and generally leads to improved ASR results (Povey, 2003; Povey et al., 2016). The MMI objective is

$$\mathcal{F}_{MMI} = \sum_{u=1}^{U} \log \frac{p_\theta(\boldsymbol{X}^u|\boldsymbol{Q}_{\boldsymbol{w}^u}) p(\boldsymbol{Q}_{\boldsymbol{w}^u})}{\sum_{\boldsymbol{w}} p_\theta(\boldsymbol{X}^u|\boldsymbol{Q}_{\boldsymbol{w}}) p(\boldsymbol{Q}_{\boldsymbol{w}})} \,, \tag{2.8}$$

where in the denominator we sum over the likelihoods for all possible state sequences $\boldsymbol{Q_w}$ weighted by their prior probabilities $p(\boldsymbol{Q_w})$. The language model is assumed to be fixed, so that only the acoustic model parameters θ need to be trained. The denominator computation is difficult because it is infeasible to enumerate all possible state sequences for non-trivial vocabulary sizes. Multiple methods have been proposed to approximate the denominator for MMI training, including n-best lists (Chow, 1990) and lattices (Valtchev et al., 1996). This makes the denominator estimation practical by limiting it to the n most likely state sequences for a given utterance or a lattice containing the most likely sequences. LF-MMI instead incorporates a complete denominator graph that can be shared across utterances. This is made possible by using a phone- instead of a word-based language model for the denominator, which is estimated on the training data.

LF-MMI models are usually trained with frame subsampling for efficiency, where only every third acoustic frame is used. The training is then repeated with different offsets until all of the data was seen. To compensate for the lower sampling rate, a special 1-state HMM topology instead of the usual 3-state topology is used, see Figure 2.3.

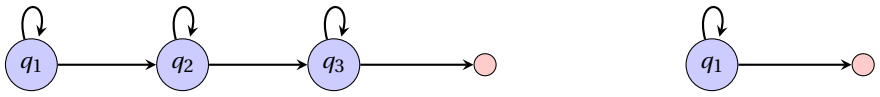


Figure 2.3: Regular 3-state and LF-MMI's 1-state HMM topology (non-emitting final states in red).

For more details and implementation notes on LF-MMI and other objective functions for acoustic model training, we refer the reader to Povey et al. (2016) and Hadian et al. (2018b).

Over the course of this thesis work, end-to-end ASR approaches have become competitive with hybrid HMM/DNN training, especially on large scale datasets. There is no single definition for end-to-end ASR, but in general these approaches simplify the training and inference pipeline by combining some or all of the feature extractor, acoustic and language models, and decoder into a single model. They also reduce the need for external resources, such as pronunciation lexicons, and for prior training of HMM/GMM models for time alignment and decision tree building. Some recent studies have also explored end-to-end approaches for pathological ASR (Hernandez et al., 2022; Hu et al., 2022; Wang et al., 2022; Yue et al., 2022a,b). In this thesis, we only consider hybrid HMM/DNN acoustic models trained with LF-MMI on MFCC features. However, LF-MMI has integrated ideas from connectionist temporal classification (CTC) training, including the 1-state HMM topology and frame subsampling. An end-to-end version of LF-MMI has also been proposed that does not depend on prior HMM/GMM training or context-dependency trees (Hadian et al., 2018b,a).

2.2 Dysarthric speech recognition

There are three main challenges for the automatic recognition of dysarthric speech. First, dysarthric speech differs from typical speech in a number of aspects, depending on the patient and their individual pathology. These in particular include a lower speaking rate, more dysfluencies, and different, less distinct and unsteady articulation (Duffy, 2012). ASR systems, which are usually trained mostly on typical speech, need to be able to model these differences.

Second, dysarthric speech itself has a lot of variability. Speech characteristics can vary significantly between different dysarthric speakers because of different individual symptoms. The speech of one speaker can also change over time because of medication, therapy, or surgery (Tykalová et al., 2015; Tripoliti et al., 2011).

Third, the large variety of different pathologies and their individual manifestations in patients results in a lack of training data with sufficient coverage. Recording speech for an extended period of time can also be strenuous for patients and existing dysarthric speech databases are therefore relatively small. Furthermore, many existing dysarthric speech corpora were not mainly recorded for the purpose of training ASR systems, but rather for dysarthria detection and assessment. The recording prompts are therefore not always representative of real-life ASR applications and recordings were carried out in controlled settings instead of naturalistic home environments. For example, participants are often asked to read phonetically complex words to assess their articulation and pronunciation, whereas in practice they might choose a simpler expression with the same meaning.

Approaches to dysarthric ASR can be grouped into three categories that try to deal with the unique characteristics and the scarcity of dysarthric speech data in different ways: model adaptation, feature adaptation, and data augmentation.

2.2.1 Acoustic model adaptation

One approach is to adapt acoustic models that were trained on potentially much larger quantities of typical speech or to modify the model training procedure to be more suited to recognising dysarthric speech.

Several studies have evaluated generic adaptation techniques for HMM/GMM acoustic models, including maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995), constrained MLLR (Gales, 1998), and maximum a posteriori (MAP) (Lee and Gauvain, 1993) to adapt speaker-independent (SI) models to a target dysarthric speaker (Sharma and Hasegawa-Johnson, 2010; Mengistu and Rudzicz, 2011a; Christensen et al., 2012, 2013; Mustafa et al., 2014). Mengistu and Rudzicz (2011a) have further added manually compiled speaker-specific pronunciation lexicons to explicitly handle lexical variability. Similarly, common adaptation strategies for hybrid HMM/DNN acoustic models, including i-vectors (Saon et al., 2013), learning hidden unit contributions (LHUC) (Swietojanski and Renals, 2014) and parameterised

activation functions (Zhang and Woodland, 2016), have been applied to dysarthric ASR (Yu et al., 2018; Liu et al., 2021). Wang et al. (2021) have also explored meta-learning to adapt DNN acoustic models to unseen dysarthric speakers.

It is also common to include typical speech data directly when training acoustic models for dysarthric speech. This leads to better ASR results than training on dysarthric speech alone. Yılmaz et al. (2016) have included typical speech from different varieties of Dutch to better handle phonetic variability.

Apart from adapting acoustic models trained on typical speech, the training process can be adjusted in other ways to better handle the characteristics of dysarthric speech. As many speech disorders affect the movement of articulators in the vocal tract, integrating articulatory information has been found to be beneficial (Rudzicz, 2011; Hahm et al., 2015; Yılmaz et al., 2018; Xiong et al., 2018; Yue et al., 2022c; Hu et al., 2022). Modelling the articulators helps to model the resulting acoustic changes. In this thesis, we train on acoustic features only, but the presented methods could also be combined with articulatory features.

2.2.2 Feature adaptation

Other works investigated transforming pathological speech to be more similar to typical speech, for example with speech enhancement methods (Bhat et al., 2018) or by adjusting speech tempo (Xiong et al., 2019). In this way, no particular adjustments have to be made to the acoustic model and a generic model for typical speech can be used to also recognise dysarthric speech. Prananta et al. (2022) investigated GAN-based VC to convert dysarthric to typical speech in order to improve dysarthric speech recognition. They also compared VC with simpler time-stretching methods and found that these work just as well.

2.2.3 Data augmentation

Data augmentation describes any method that allows to create additional training data, either by transforming the original speech or by adapting external speech data to be similar to the target data. It is commonly used in many low-resource ASR settings. Data augmentation is also relevant for other pathological speech processing tasks, such as dysarthric speech detection (Jiao et al., 2018).

Many methods described in the previous section on feature adaptation can also be employed for data augmentation by applying them in reverse. Instead of adapting dysarthric speech to be more similar to typical speech for use in generic ASR systems, the opposite is done and additional dysarthric training data is created from typical speech. Typical speech data is abundant, so this allows to generate a large amount of data, while the original dysarthric speech can still be included in acoustic model training as well.

Data transformations

The simplest forms of data augmentation are basic transformations of the data that often do not specifically model the speakers or their pathology. These include speed perturbation (Ko et al., 2015), where copies of the training data with slightly perturbed speed are created, and SpecAugment (Park et al., 2019), where random chunks of either time or frequency information are masked for each utterance. SpecAugment is particularly effective for large scale datasets (Park et al., 2020) and therefore less suited for dysarthric ASR.

Other transformations specifically aim to create artificial dysarthric speech from typical speech, for example by adjusting the speed (not preserving the pitch) or tempo (preserving the pitch) to match the speaking rates of dysarthric speakers (Vachhani et al., 2018; Xiong et al., 2019).

Voice conversion

VC is the task of converting a speech signal with the voice of one speaker to that of another speaker while preserving the linguistic content as illustrated in Figure 2.4. It is an attractive method to improve dysarthric speech recognition because it allows to make dysarthric speech more similar to typical speech. Alternatively, typical speech may be converted to dysarthric speech to generate additional training data. Recent deep learning-based VC systems typically use GANs (Goodfellow et al., 2014) or encoder-decoder architectures (Sisman et al., 2021). There are parallel VC approaches that require source and target speech samples with matching linguistic content and non-parallel ones that place no restrictions on the contents of the training data.

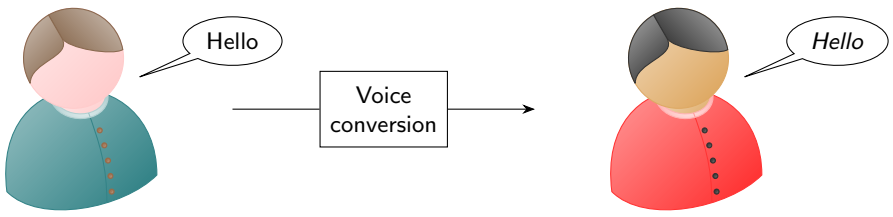


Figure 2.4: Illustration of the voice conversion task.

Both GAN-based (Jin et al., 2021, 2022a,b) and sequence-to-sequence typical-to-dysarthric VC (Harvill et al., 2021) have been used as data augmentation for dysarthric ASR in a number of studies.

Illa et al. (2021) have studied VC between dysarthric speakers. Illa et al. (2021), Halpern et al. (2021) and Huang et al. (2022) have also compared the generated dysarthric speech outputs with real dysarthric speech in both subjective and objective tests.

Speech synthesis

Speech synthesis, or TTS, refers to generating a speech signal from a sequence of words. Similar to voice conversion, it allows to create additional training data to improve dysarthric speech recognition.

But unlike VC, which requires existing recordings of the target utterances, we can synthesise speech for arbitrary sentences and therefore quickly adapt an ASR system to new commands and domains and a single model can handle any number of speakers. TTS-based data augmentation has already been applied to ASR for low-resource languages and children's speech (Kadyan et al., 2021). ASR and TTS are also naturally linked, corresponding to speech perception and speech production, and Tjandra et al. (2017, 2020) have proposed joint training to emulate the human speech chain and incorporate additional unlabelled data.

Soleymanpour et al. (2022) have introduced a dysarthric TTS system that allows to explicitly model and generate speech of different severity levels. They showed that it is effective as data augmentation in a dysarthric ASR system.

2.3 Datasets

In this section, we present the datasets used throughout this thesis. While our main focus is on dysarthria, we also consider children's speech and non-native speech as other forms of atypical speech. We describe the corresponding language models and pronunciations lexicons in the relevant later chapters.

2.3.1 Torgo

The Torgo corpus of dysarthric speech (Rudzicz et al., 2012b) contains about 15 hours of recordings from 15 Canadian English speakers (six female, nine male) made with an array and a headset microphone. There are one ALS and seven cerebral palsy patients aged 16–50 with mostly severe dysarthria (6 hours of speech in total) and seven age-matched control speakers without any speech impairment (9 hours of speech in total). Each dysarthric speaker was assessed by a speech and language pathologist according to the Frenchay Dysarthria Assessment (Enderby, 1980). It comprises evaluations in different categories, including respiration, tongue, lips, and intelligibility, each of which are rated on a 9-point scale.

Participants were asked to complete different recordings tasks, such as word and sentence reading, picture descriptions, sustained vowels, and diadochokinetic tasks. For ASR training we include only the isolated word and sentence recordings in line with previous works (España-Bonet and Fonollosa, 2016). We further discard utterances that have no transcriptions or that are too short to contain any speech.

Table 2.2: Torgo corpus statistics.

Total utterances	16394
Total unique utterances	971
Total multi-word utterances	4161
Total unique multi-word utterances	356

Table 2.2 provides statistics on the utterances that we include for ASR training. It shows that the number of unique utterances is small, meaning that many are repeated within and across speakers (Yue et al., 2020). About 75% of utterances consist of isolated words, among which are many minimal pairs, such as *rate* and *raid* without context that would disambiguate them. In fact, for 88% of isolated words there is at least one other word with a pronunciation within an edit distance of one. The average closest edit distance is 1.16. This makes the corpus very useful for automatic assessment of speech intelligibility and similar tasks, but more challenging for ASR. Even for speakers without any speech disorders correctly distinguishing the minimal pairs is expected to be difficult.

In addition to the speech recordings, Torgo also contains time-aligned articulatory movement data recorded with a 3-dimensional electromagnetic articulograph. This allows ASR systems to incorporate articulatory data, but in this work we only make use of the acoustic data.

2.3.2 UA-Speech

The UA-Speech corpus (Kim et al., 2008) consists of recordings of isolated words made with a 7-channel microphone array from 15 adult dysarthric speakers (four female, 11 male) with cerebral palsy (about 40 hours in total) and 13 control speakers (four female, nine male) without any speech impairment (about 30 hours in total). All are speakers of American English.

We include the data from all microphone channels and use the re-segmented version of the corpus from Xiong et al. (2019) where excessive portions of silence have been removed via forced alignment with an HMM/GMM ASR system. The set of words is divided into three distinct blocks, two of which (Block 1 and 3) are commonly used as the training set, while models are evaluated on Block 2. 155 words are common across all blocks and then each block contains a unique set of another 100 words, for a total of 455 different words in the corpus. The vocabulary includes numbers (e.g., “thirty-five”), computer commands (e.g., “delete”) and lists of common (e.g., “with”) and uncommon (e.g., “beleaguering”) words.

The intelligibility of the dysarthric speakers was assessed by human listeners with a word transcription task and is reported in percent. We group our ASR results by speaker severity based on these intelligibility ratings in the same way as previous works (Xiong et al., 2019; Hernandez et al., 2022), see Table 2.3.

Table 2.3: Severity and intelligibility of UA-Speech dysarthric speakers.

Severity	Speaker	Intelligibility (%)
Severe	M04	2
	F03	6
	M12	7.4
	M01	15
Moderate-severe	M07	28
	F02	29
	M16	43
Moderate	M05	58
	M11	62
	F04	62
Mild	M09	86
	M14	90.4
	M10	93
	M08	93
	F05	95

We note that both Torgo and UA-Speech contain significant amounts of background noise that further affect ASR results (Schu et al., 2023).

2.3.3 PF-STAR

We use the PF-STAR corpus (Batliner et al., 2005) for experiments on children's speech recognition. It recorded a mix of sentences and isolated words in British English from 152 children aged 4 to 14 years with two different microphones. The training set contains 15 hours of speech from 80 speakers with the data from both microphones.

2.3.4 Non-native children's speech (FBK and ETS)

We participated in the Interspeech 2021 Shared Task on ASR for Non-Native Children's Speech (Gretter et al., 2021) and used their provided data. For English, 50 hours of recordings from an English proficiency test of 800 speakers aged 11 and above were provided by ETS. The German data consists of 5 hours of recordings from a proficiency test of 300 Italian students collected by FBK (Gretter et al., 2020). Additional untranscribed German data was available for semi-supervised training that we did not use. Separate development and test sets were also given for both languages.

2.3.5 Librispeech

It is important to avoid excessive hyperparameter tuning on the relatively small dysarthric speech corpora to avoid overfitting because they do not have designated validation sets. Unless otherwise mentioned, we therefore base our acoustic models on the well-tuned Kaldi recipes for the 5-hour subset of the Librispeech corpus (Panayotov et al., 2015) of read speech from audiobooks.[1]

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/mini_librispeech

3 Sequence-discriminative training for dysarthric speech recognition

3.1 Introduction

Until recently, most works on dysarthric speech recognition have been in the framework of maximum likelihood-trained HMM/GMM models or hybrid HMM/DNN models trained with a frame-level cross-entropy (CE) objective. However, as ASR is a sequence modelling problem, recent state-of-the-art systems are increasingly trained with sequence-discriminative loss functions, especially LF-MMI (Povey et al., 2016) and CTC (Graves et al., 2006). The use of such sequence-discriminative criteria has not been sufficiently explored in the context of pathological speech yet. LF-MMI has previously been applied to dysarthric speech (Xiong et al., 2019), but its performance in comparison with other methods has not yet been analysed in detail.

Multiple now common techniques employed for performance or efficiency reasons in LF-MMI training and other state-of-the-art models, such as frame subsampling (Sak et al., 2015) and speed perturbation (Ko et al., 2015), potentially also give performance benefits especially on dysarthric speech. For frame subsampling, only every third frame is preserved during training and decoding for a substantial speedup. At training time, this sampling is repeated with different offsets, so that eventually the model still sees every frame. Similarly, for dysarthric speech recognition it was suggested to increase the frame shift of dysarthric speakers during feature extraction to compensate for their lower speaking rates (España-Bonet and Fonollosa, 2016). Speed perturbation augments the training data with multiple (usually 2) copies of itself with slightly modified speed to make models more robust to different speaking rates and to increase the amount of training data, which is crucial for neural network training on small corpora. It could thus also help with the much larger speaking rate variability found in dysarthric speech. We therefore focus our analysis on these techniques.

We evaluate our systems on the Torgo and UA-Speech corpora. Unlike previous works, we report the Torgo results separately for isolated and multi-word utterances to make them more informative. We show that time-delay neural network (TDNN) acoustic models trained with the LF-MMI objective give state-of-the-art results and especially reduce the number of

insertion errors. ASR systems commonly insert many spurious words when encountering dysarthric speech (Moore et al., 2019) because it is often much slower than the speech they are typically trained on. These LF-MMI models serve as baselines throughout this thesis.

The remainder of this chapter is organised as follows. Section 3.2 describes our experimental setup for training ASR systems and the evaluation protocol. In Section 3.3, we present our results and analyse the strong performance of LF-MMI systems. Section 3.4 summarises the main contributions of this chapter.

3.2 Experimental setup

3.2.1 Systems

In Figure 3.1 we provide a schematic overview of the acoustic models trained in this thesis and how we refer to them. We train SI models, where we train a single model on either all dysarthric, all control, or both sets of speakers. We note that for UA-Speech these nonetheless include the training data of the target dysarthric speakers from blocks 1 and 3. For completeness, we also compare these with fully SI models trained on all data except that of the target speaker. There is no speaker overlap between training and test data for Torgo because of a cross-validation evaluation protocol, described in more detail in Section 3.2.2.

Furthermore, we train two sets of speaker-dependent (SD) models, where separate acoustic models are trained on only the training data of the target dysarthric speaker from UA-Speech or that speaker's data and all control speech (SD + Control).

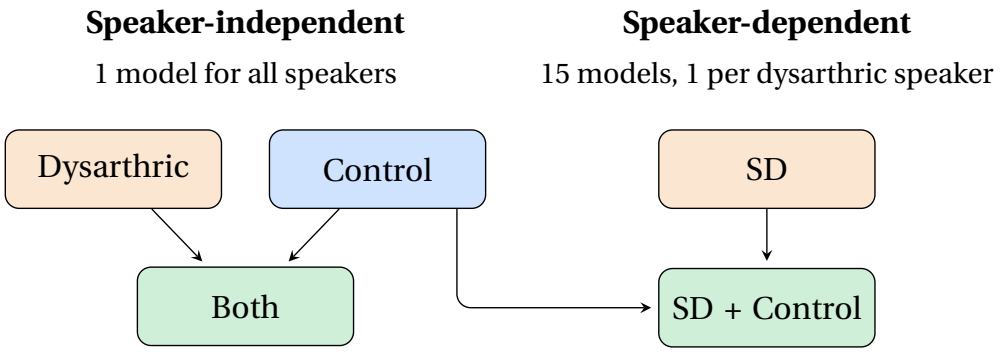


Figure 3.1: Overview of speaker-independent and speaker-dependent (SD) acoustic models trained in this thesis.

HMM/GMM

We use the open-source Kaldi speech recognition toolkit (Povey et al., 2011) for all our ASR experiments. We followed the typical pipeline of successively training monophone, triphone, speaker adaptive training (SAT), and subspace GMM (SGMM) (Povey et al., 2010) baseline

models on 39-dimensional MFCC+Δ+ΔΔ features. We use the hyperparameters and provided Kaldi recipe[1] of España-Bonet and Fonollosa (2016) for the Torgo corpus and the recipe[2] of Xiong et al. (2019) for UA-Speech as a basis.

We use the CMU pronunciation dictionary and its Arpabet phone set for both corpora, generating missing pronunciations with the Phonetisaurus grapheme-to-phoneme conversion toolkit (Novak et al., 2016) and subsequent manual verification. In contrast to established procedures for typical speech, we also chose to model phones independent of their position in words as suggested by Joy and Umesh (2018) because of data sparsity and because the lower speaking rates of dysarthric speech lead to reduced coarticulation effects.

HMM/DNN

The main system for Torgo that we analyse below is a 13-layer factorised TDNN model (Povey et al., 2018) trained with the sequence-discriminative LF-MMI objective function. We use regular LF-MMI that requires alignments from a previously trained GMM model. Throughout this thesis, we use the SAT model for this. For comparison, we also trained a 9-layer TDNN-LSTM model with a conventional frame-wise CE objective. As is the default in Kaldi, we trained the HMM/DNN models on speed perturbed data for which the training data is augmented by perturbed versions at 0.9 and 1.1 times the original speed (Ko et al., 2015).

For UA-Speech, we train a LF-MMI model with six convolutional neural network (CNN) layers followed by nine factorised TDNN layers, also using speed perturbation, adapted from Xiong et al. (2019). We also compare this with training the same model with a frame-wise CE objective function instead.

3.2.2 Evaluation protocol

Because Torgo contains only eight dysarthric speakers and their degree of dysarthria varies significantly, we maintain the leave-one-out cross-validation training procedure of España-Bonet and Fonollosa (2016), where each of the 15 speakers is evaluated separately and models are trained on the remaining 14 speakers.

Unlike previous works, we split the evaluation of isolated- and multi-word utterances by treating the two tasks separately. Otherwise the results would be less informative because of the different challenges in these two tasks. Most prior research on dysarthric speech recognition has focused on isolated words because of the lack of datasets that include continuous speech. However, we do not see this as a limitation. Speaking can require a significant effort from severely dysarthric speakers and to maximise communication efficiency they might choose to use shorter utterances. For example, the homeService (Nicolao et al., 2016) and EasyCall (Turrisi et al., 2021) corpora were recorded in realistic home environments and contain simple

[1] https://github.com/cristinae/ASRdys
[2] https://github.com/ffxiong/uaspeech/tree/master/s5_segment

commands of one or two words, such as *"Volume up"*. Most other pathological speech corpora are not recorded specifically for ASR, but for speech assessment purposes, which explains why the sentences in the Torgo corpus are often long and unnatural.

The language models (LMs) are different for the two evaluation tasks. For isolated word recognition it is a unigram model containing all 615 possible words, which may be preceded or followed by silence. In Section 3.3 we also evaluate the effect of constraining the decoding grammar so that the output is always a single word. For sentences we use a bigram LM that is trained on all the sentence data. In both cases we train the LMs on the data of all speakers, they thus also include that of the test speaker. This is impossible to avoid because there is very high text overlap between speakers as explained in Section 2.3.1 and in this way we focus on improving the acoustic model. Improvements on the LM side could only be obtained with LMs trained on large external corpora because the Torgo corpus is so small. This has been studied in more detail by Yue et al. (2020). The language model weight for decoding in each experiment was set to the average of the best values obtained for each control speaker.[3]

For UA-Speech, we simply follow the standard protocol of training on blocks 1 and 3 of the data and evaluating on block 2. We use a decoding grammar that only contains the possible output words and restricts the output to a single word, similar to Xiong et al. (2019).

3.3 Results and analysis

Table 3.1 shows the results for evaluating the Torgo baseline systems described in Section 3.2.1 separately on isolated-word and multi-word utterances. The WERs are aggregated across dysarthric and control speakers for readability, but there can be substantial variation between individual speakers depending on their severity as illustrated in Figure 3.3 and as we will show in more detailed breakdowns for UA-Speech.

As predicted in Section 2.3.1, WERs on the isolated word task are high even for the control speakers because of the inherent challenge in distinguishing minimal pairs without further context. Figure 3.2 highlights that most mistakes in the ASR output for the isolated word task on control speech are close calls with a Levenshtein distance of one or two phonemes to the reference transcript. The majority of errors on dysarthric speech, however, involve more phoneme changes.

On the other hand, the sentences are recognised with only very few errors for the control group and mildly dysarthric speakers because the strong LM renders this task quite easy. Despite this advantage, WERs for moderate to severely dysarthric speakers are much higher in this case.

[3] In Kaldi it is common to perform a grid search for language model weight and word insertion penalty at decoding time even on test data because differences are often small, but with the cross-validation setup on the Torgo corpus it is important to avoid tuning any parameters on a specific dysarthric speaker's data because the impact might be much larger.

Table 3.1: WER on Torgo for SGMM, CE, and LF-MMI systems, averaged for dysarthric and control speakers, respectively. Every second row shows the effect of restricting the output to a single word during isolated word recognition.

	1-word	**Isolated**		**Sentences**	
	LM	Dysarthric	Control	Dysarthric	Control
SGMM	✗	56.1	19.4	41.5	4.4
	✓	47.2	18.7	–	–
CE	✗	53.6	24.6	38.0	9.3
	✓	44.9	24.0	–	–
LF-MMI	✗	49.2	24.0	25.9	7.9
	✓	43.0	22.0	–	–

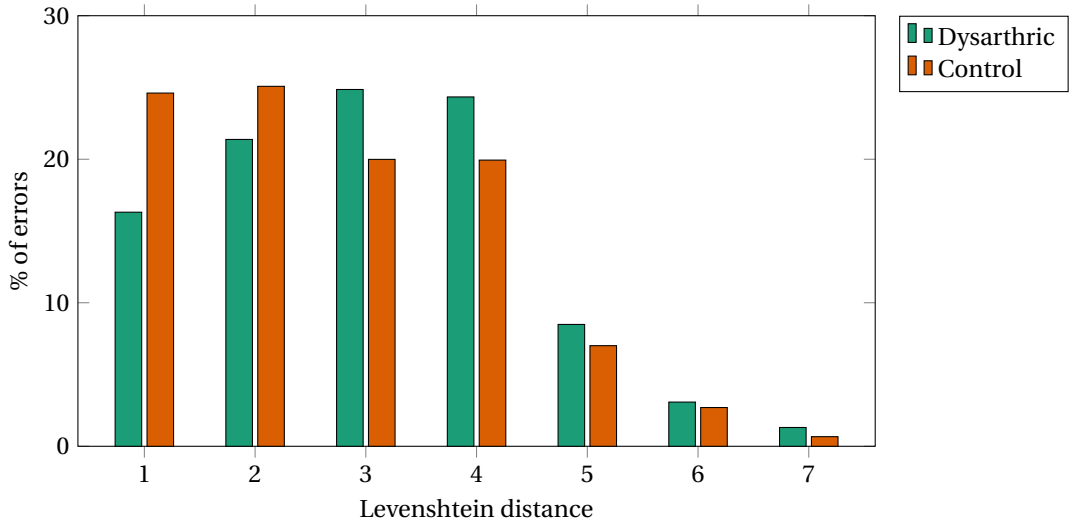


Figure 3.2: Distribution of Levenshtein distances between the phoneme sequences of the reference transcripts and the ASR errors on the Torgo isolated word task of the LF-MMI acoustic model trained on both dysarthric and control speakers.

This highlights that there is still a lot of room for improvement just on the acoustic modelling side.

LF-MMI training always helps for dysarthric speech except for one speaker compared to both SGMM and CE-based models. However, the SGMM outperforms the neural network models on the control speakers, perhaps because on such a small corpus the neural networks are more sensitive to the additional variability in the training data introduced by the dysarthric speech. Indeed, if the LF-MMI system is trained for the same number of epochs and with the same hyperparameters on the control speech only, it performs much better, with WERs of 18.3% on the isolated words and 2.9% on the sentences averaged over all control speakers.

The large improvements of LF-MMI models on the sentences are because they make much fewer insertion errors, indicating that they are better equipped to handle very low speaking rates. Figure 3.3 shows how speaking rate and WER are correlated. We approximate speaking rate information by computing mean phoneme durations from forced alignments of the training data with the methodology of Xiong et al. (2019). It can be seen that dysarthric speakers have the lowest speaking rates and also the highest WERs. There are another three mildly dysarthric speakers that have normal or even slightly shorter phoneme durations that the ASR system recognises very well.

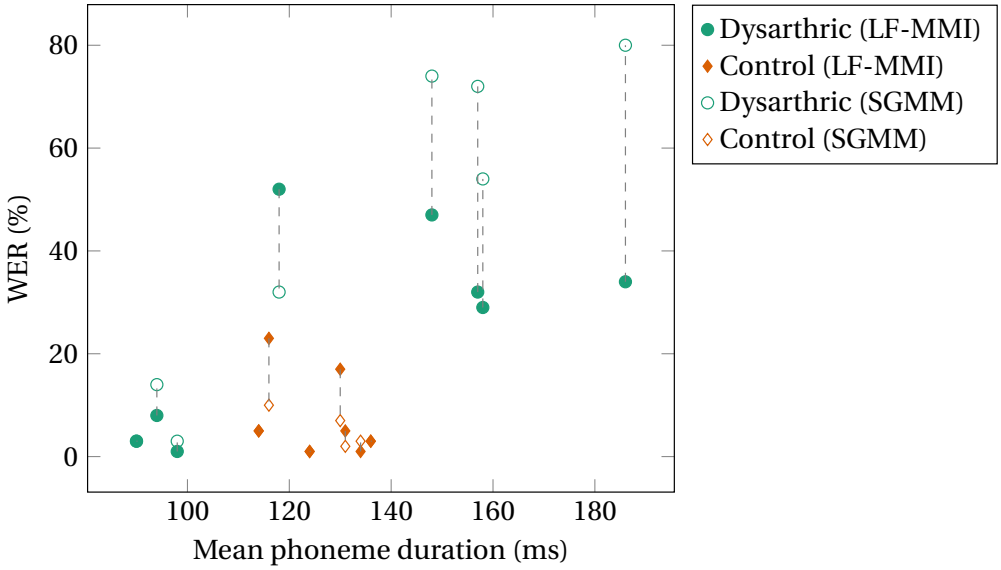


Figure 3.3: Relationship between mean phoneme duration and WER for dysarthric and control speakers. The WER results are from the LF-MMI and SGMM models on the Torgo sentence task. Mildly dysarthric speakers also achieve very low WER. Dashed lines connect results from the same speaker.

For the sake of completeness, we also evaluated all speech grouped together and with the methodology of España-Bonet and Fonollosa (2016). We substantially outperform their best results obtained with hybrid HMM/DNN systems that were the previous state of the art on this corpus.

In the following sections, we analyse the performance of LF-MMI in more detail.

3.3.1 Constrained language model

Every second row in Table 3.1 also shows the results of forcing the decoder to output only a single word for the isolated-word utterances. This consistently improves results across speakers, in particular for the most severely dysarthric ones because their very low speaking rate otherwise leads to a large number of insertion errors. This suggests that the WER on the sentence task where the number of words is not known a priori could also be reduced by

appropriately tuning the word insertion penalty during decoding for each speaker or utterance. However, this penalty would need to be set in an unsupervised manner by automatically estimating speaking rates.

3.3.2 Speed perturbation

As mentioned in Section 3.2, the training data for the hybrid HMM/DNN systems was augmented with two speed-perturbed copies. To test the effects of this we trained LF-MMI models on the original data only, but increasing the number of epochs by a factor of three to compensate for the lower amount of training data. Results, shown in Table 3.2, are overall still better than SGMMs and cross-entropy models, maintaining a big reduction in insertion errors as indicated by the sentence results. This suggests that this reduction can at least in part be attributed to the sequence-discriminative objective function.

However, the performance on control speech is better when no speed perturbation is applied. It is then on par with the SGMM results, but still worse than training on speed-perturbed control speech only as observed above. This is perhaps because applying further distortions to dysarthric speech makes the training data too variable to perform well on unimpaired speech.

Table 3.2: LF-MMI systems trained without speed perturbation still outperform SGMMs on Torgo. The isolated word results use the constrained LM.

	Speed	**Isolated**		**Sentences**	
	perturbation	Dysarthric	Control	Dysarthric	Control
SGMM	✗	47.2	18.7	41.5	4.4
LF-MMI	✓	43.0	22.0	25.9	7.9
	✗	46.4	21.4	30.2	4.2

3.3.3 Frame shift

Previous work (España-Bonet and Fonollosa, 2016) proposed to apply a frame shift of 15 ms to the dysarthric data during MFCC extraction while maintaining the usual 10 ms for the control speech to compensate for the lower speaking rates of dysarthric speakers. However, the good performance of the LF-MMI systems suggests that it might not be necessary in these models. Our results in Table 3.3 confirm that a constant frame shift of 10 ms for the entire data does not reduce performance on dysarthric speech. This is useful because the constant frame shift does not require prior knowledge about the speaker or any special processing and we maintain this throughout all following experiments in this thesis.

Table 3.3: Applying a 15 ms frame shift to dysarthric and 10 ms to control speakers compared with a constant 10 ms shift throughout on Torgo. The isolated word results use the constrained LM.

	Frame shift	**Isolated**		**Sentences**	
		Dysarthric	Control	Dysarthric	Control
LF-MMI	15/10 ms	43.0	22.0	25.9	7.9
	10 ms	42.9	22.5	25.9	8.1

3.3.4 UA-Speech

Results for the baseline systems on UA-Speech can be found in Table 3.4. We compare SD and SI systems and training on all dysarthric, all control, or both sets of speakers for SGMM, CE, and LF-MMI models. We show aggregate results for each severity group, and for the dysarthric and control speakers overall. For reference, we also include directly comparable results of Xiong et al. (2020) that were obtained from LF-MMI acoustic models with a similar Kaldi recipe as in our work.

We observe the same pattern of strong improvements of LF-MMI over SGMMs as on Torgo. However, differences between CE and LF-MMI are relatively small here. This is probably because of the short utterances of isolated words in UA-Speech, whereas LF-MMI showed a clearer benefit over CE on the Torgo sentence data. Nonetheless, LF-MMI excels on more severely dysarthric speech. Again, we see that SGMMs outperform the other methods on control and mildly dysarthric speech.

We also experimented with i-vectors for speaker adaptive neural network training, but found they only provide minor benefits if at all. For simplicity, all models for Torgo and UA-Speech in this thesis are therefore trained without i-vectors. Findings from Liu et al. (2021) indicate that LHUC can lead to bigger improvements than i-vectors on dysarthric speech.

3.4 Summary

The goal of this chapter was to evaluate whether state-of-the-art LF-MMI acoustic model training could also be applied to dysarthric ASR. We demonstrated that it indeed yields strong results on such small and atypical datasets, in particular for more severely dysarthric speakers. On mildly dysarthric and control speakers, SGMMs remained competitive. Our results were a new state of the art on the Torgo corpus that can serve as strong baselines for further research. When analysing these improvements we found that especially insertion errors are reduced, which are otherwise very frequent due to the low speaking rates of dysarthric speakers. Contributing factors to this are the frame subsampling of LF-MMI, data augmentation with speed perturbed speech and the sequence-discriminative objective function itself. Further analysis is required to determine the importance of each of these factors. While hybrid

Table 3.4: Word error rates (WER) on the UA-Speech corpus for each group of speakers. Our best results in each column are highlighted in bold.

Systems	**Dysarthric**					**Control**
	Severe	Mod.-Severe	Moderate	Mild	Overall	
SGMM						
SD	86.4	60.5	74.7	41.8	62.5	-
SI	91.0	75.8	56.6	29.9	58.7	-
Dysarthric	77.1	42.6	38.1	17.6	40.5	20.4
Control	96.1	83.4	62.0	20.8	59.3	10.2
Both	78.0	43.8	34.9	**14.2**	39.0	**9.8**
CE						
SD	66.6	36.9	34.5	18.9	36.6	-
SI	92.5	77.8	56.8	33.3	60.7	-
Dysarthric	64.1	33.6	27.9	16.7	33.2	23.9
Control	95.7	82.3	65.0	26.8	61.7	10.8
Both	66.0	36.3	27.7	16.6	34.2	12.3
LF-MMI						
SD	70.3	42.7	38.2	24.0	41.3	-
SD + Control	65.5	32.8	**25.8**	15.7	**32.6**	-
SI	89.0	67.7	49.7	31.3	55.7	-
Dysarthric	**62.4**	**32.2**	29.2	19.0	34.0	24.0
Control	96.2	74.5	55.1	23.2	56.9	14.0
Both	62.8	35.7	28.7	17.7	33.9	15.5
LF-MMI from Xiong et al. (2020)						
SD	70.9	33.7	31.4	14.6	34.8	-
SD + Control	67.1	34.4	25.7	13.3	32.4	-
SI	90.8	71.3	51.9	32.4	57.7	-
Dysarthric	63.0	30.9	28.2	18.9	33.3	-
Control	97.2	78.5	56.4	19.3	56.8	-

HMM/DNN systems reduce the number of errors on dysarthric speech, we observed that they do not work as well for control speakers as systems trained only on control speech or a traditional HMM/GMM system. In the next chapter we propose a solution for this in order to improve speech recognition for everyone.

4 Model combination for dysarthric speech recognition

4.1 Introduction

In this chapter we propose to overcome the acoustic- and pronunciation-level mismatch between dysarthric and typical speech through a unified acoustic model combination approach. While we have shown in the previous chapter that adding control speech to the training data leads to improved results on dysarthric ASR, we argue that simply adding even more of it cannot be sufficient because data scarcity is not the only reason for poor performance of ASR systems on dysarthric speech. The specific differences of dysarthric speech also need to be accounted for.

The variation in dysarthric speech with respect to control speech lies at acoustic level as well as at pronunciation level. Dealing with those two variations separately is not a trivial task for two main reasons. First, the lexicon is based on control speech, i.e. highly intelligible, typically native speech. Second, the set of acoustic units is determined by phonotactic constraints enforced by the lexicon, even though dysarthric speech may be unintelligible and the phoneme sequences not clearly identifiable. We further illustrate these issues with our proposed analysis framework in Chapter 5.

One possible way to handle this challenge is to manually update the lexicon with pronunciation variants (Mengistu and Rudzicz, 2011a). This is still challenging as dysarthric speech intelligibility varies with the level of severity, with severely dysarthric speech being almost unintelligible, so that it would be difficult even for human expert listeners to transcribe the phone or sound sequence in a consistent manner. Another option is to automatically build speaker-specific pronunciation dictionaries based on mispronunciation analysis (Sriranjani et al., 2015). In this work, however, we are interested in speaker-independent methods. Also, as noted earlier, even speech of the same speaker can vary due to medical condition or therapy.

Saraçlar and Khudanpur (2004) have shown that pronunciation variation in conversational speech with respect to dictionary forms can be handled implicitly by ASR systems through informed clustering of HMM states. Furthermore, pronunciation models that dynamically

adapt based on contextual information, e.g. surrounding words or speaking rate, can help in handling pronunciation variation of spontaneous speech (Fosler-Lussier, 1999a,b). We build upon these two aspects to propose two approaches to implicitly handle acoustic- and pronunciation-level mismatch between control speech and dysarthric speech: combination of models trained on different speaker groups and combination of models trained with different subword units.

Previously, we assumed the pronunciation lexicon to be based on phonemes, but graphemes are a popular alternative choice of subword unit. For low-resource languages, they obviate the need for manually created phoneme lexicons (Gales et al., 2015). Alternatively, grapheme-based methods may also enable to build pronunciation lexicons in resource-constrained scenarios in the first place (Rasipuram and Doss, 2012; Rasipuram et al., 2013). Recent end-to-end ASR methods are trained almost exclusively with graphemes both for convenience and because they often perform better than phonemes (Irie et al., 2019).

Particularly languages with a deep orthography, like English, do not have a clear correspondence between graphemes and phonemes. This gives another motivation to use graphemic units in ASR systems for atypical speech, namely relaxing constraints imposed by a phoneme-based lexicon and allowing the acoustic model to learn how to map phonemes to graphemes on its own (Tejedor et al., 2008).

We propose to train separate acoustic models on different speaker subsets of the data and combine them by dynamic acoustic model selection during decoding. Thus, each model is specialised for the acoustic characteristics of its training speakers, but with model combination we can still recognise a variety of speech conditions without requiring prior information about the speaker. We then further extend this approach to also handle pronunciation variation by combining models trained with different acoustic subword units like phonemes and graphemes and evaluate this also on children's and non-native children's speech. Our proposed approach is also related to multi-stream ASR (Hermansky, 2013), which integrates multiple parallel information streams and where the model falls back to more reliable streams when multiple streams present conflicting information.

4.2 Proposed approach

Mathematically, we consider an HMM-based ASR system to estimate the joint probability of a word hypothesis $\boldsymbol{w}^u = (w_1, \ldots, w_M)$ and the observed sequence of acoustic feature vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ by summing over all possible sequences $\boldsymbol{Q}$ of HMM states. These are commonly decomposed into an acoustic and a language model (see also Section 2.1):

$$p(\boldsymbol{w}^u, \boldsymbol{X}) = \sum_{\boldsymbol{q} \in \boldsymbol{Q}} p(\boldsymbol{w}^u, \boldsymbol{X}, \boldsymbol{q}) \tag{4.1}$$

$$= \sum_{\boldsymbol{q} \in \boldsymbol{Q}} p(\boldsymbol{X}, \boldsymbol{q} | \boldsymbol{w}^u) p(\boldsymbol{w}^u) \ . \tag{4.2}$$

As elucidated in Rasipuram and Magimai.-Doss (2015), the acoustic model likelihood estimation $p(\boldsymbol{X}, \boldsymbol{q}|\boldsymbol{w}^u)$ combines the HMM state emission and transition probabilities at each of T time steps, which, after i.i.d and first order Markov assumption, can be written as:

$$p(\boldsymbol{X}, \boldsymbol{q}|\boldsymbol{w}^u) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|q_t)\ P(q_t|q_{t-1}) \tag{4.3}$$

$$= \prod_{t=1}^{T} \left(\sum_{d=1}^{D} p(\boldsymbol{x}_t|a^d)\ P(a^d|q_t) \right) \cdot P(q_t|q_{t-1}) \ , \tag{4.4}$$

where $\{a^d\}_{d=1}^{D}$ is the set of D acoustic units, whose likelihood $p(\boldsymbol{x}_t|a^d)$ can be estimated with a GMM or neural network. In this work we follow the common strategy of using clustered context-dependent states, or *senones*, as the acoustic units, where $p(a^d|q_t)$ is a Kronecker delta distribution determined by a decision tree that maps HMM states modelling context-dependent subword units to senones in a one-to-one manner. A more complex, probabilistic mapping from HMM states to acoustic units would also be possible though and is employed in the Kullback-Leibler divergence based hidden Markov model (KL-HMM) approach (Aradilla et al., 2007, 2008), which has also been successfully applied to dysarthric ASR (Kim et al., 2016) on a Korean speech corpus.

4.2.1 Model combination

To handle acoustic and lexical variation, we train acoustic models on different speaker subsets of the data and combine them by dynamic acoustic model selection during decoding. Formally, this can be expressed as

$$p(\boldsymbol{w}^u, \boldsymbol{X}) = \max_j p(\boldsymbol{w}^{j,k}, \boldsymbol{X}) \ , \tag{4.5}$$

where $j \in \{\text{control}, \text{dysarthric}, \text{both}\}$ denotes the subset of data used for training the acoustic model. This yields different estimators for $p(\boldsymbol{x}_t|a^d)$ and $P(a^d|q_t)$, so we can write the acoustic model likelihood estimation as

$$p(\boldsymbol{X}, \boldsymbol{q}^j | \boldsymbol{w}^{k,j}) = \prod_{t=1}^{T} \left(\sum_{d=1}^{D^j} p(\boldsymbol{x}_t | a^{d,j}) \; P(a^{d,j} | q_t) \right) P(q_t | q_{t-1}) \;\; , \tag{4.6}$$

In this way, the acoustic characteristics of each group of speakers can be modelled separately. The decision tree clustering also results in distinct sets of senones for each model, capturing lexical variability to a certain degree.

4.2.2 Alternative pronunciation models

In our alternate pronunciation models approach we build upon the understanding that grapheme units could relax the constraints imposed by phoneme lexicons on acoustic models for dysarthric speech and train two separate acoustic models with phoneme and grapheme lexicons, which are then selected dynamically during decoding (as in Equation (4.5)). We can formally express the acoustic model likelihood estimation in this case as

$$p(\boldsymbol{X}, \boldsymbol{q}^j | \boldsymbol{w}^{k,j}) = \prod_{t=1}^{T} \left(\sum_{d=1}^{D^j} p(\boldsymbol{x}_t | a^{d,j}) \; P(a^{d,j} | q_t^j) \right) P(q_t^j | q_{t-1}^j) \;\; , \tag{4.7}$$

where $j \in \{\text{phoneme lexicon}, \text{grapheme lexicon}\}$.

4.3 Experimental setup

Dysarthric speech

We use the UA-Speech and Torgo dysarthric speech corpora for our experiments in this chapter to ensure that the results are not specific to a single dataset. The baseline systems are the SGMM and LF-MMI models from the previous chapter, trained on only dysarthric, only control, or both sets of speakers.

We combine the trained acoustic models by computing the union of the decoding lattices with subsequent minimum Bayes risk (MBR) decoding as it is implemented in Kaldi through the `lattice-combine` binary (Xu et al., 2011). During lattice combination, the total probability of each path is normalised, which allows to combine any two models even when they are trained with different acoustic unit sets as in our case. Recogniser output voting error reduction (ROVER) (Fiscus, 1997) is another common method to combine decoding hypotheses from multiple ASR systems and has been used in a similar way to combine different pronunciation models (Fosler-Lussier, 1999a).

On UA-Speech, we also experiment with alternative pronunciation models, where we additionally train systems with grapheme units instead of phonemes as before. We create such a graphemic lexicon for the words in the UA-Speech corpus by representing each letter as

a grapheme. We then proceed to train GMM and LF-MMI models exactly like we did for phonemes, including both dysarthric and control speech in the training data. We then also combine the outputs from the phoneme and grapheme system with lattice combination and subsequent MBR decoding as above.

Children's speech

We further evaluate the alternative pronunciation model approach on children's and non-native children's speech. For this we use the PF-STAR children's speech corpus and German and English data from the Interspeech 2021 Shared Task on ASR for Non-Native Children's Speech (Gretter et al., 2021).

For the PF-STAR baseline, we first train a GMM system from an existing recipe (Dubagunta et al., 2019), using the BEEP pronunciation lexicon.[1] With the GMM alignments, we then train a LF-MMI model with six CNN layers followed by nine factorised TDNN layers. We use i-vectors, speed perturbation and SpecAugment data augmentation (Park et al., 2019). We use the two language models of Dubagunta et al. (2019), which were trained on the PF-STAR training set transcriptions and additional data from the MGB challenge (Bell et al., 2015) and then interpolated. The `eval/adapt` portion of the corpus serves as a development set. We report the results from the two microphones (*test-A* and *test-B*) separately.

Baseline LF-MMI Kaldi recipes for the shared task were provided for both languages and are described in more detail by Gretter et al. (2021). We re-train these baselines ourselves, but do not make any modifications. The English language model is trained only on the training transcriptions, the German one on the provided additional text data.

For both of these datasets we also create grapheme lexicons as above and train grapheme-based models in the same manner.

4.4 Results and discussion

4.4.1 Baselines

Table 4.1 shows the WERs of the Torgo baseline models on the isolated word and sentence recognition tasks. As we have shown in the previous chapter, ASR systems trained with a LF-MMI loss function bring significant improvements over traditional SGMM based systems for most dysarthric speakers. For best results, control speech should be added to the training data. However, LF-MMI models then perform worse on control speakers than GMMs or a system trained on control speakers alone for both tasks. This might not be desirable, for example when developing general purpose ASR systems where the target audience is more likely not known a priori.

[1] http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

Table 4.1: WER results on Torgo, averaged over dysarthric and control speakers, respectively. The acoustic models were trained either on both dysarthric and control speakers, or only one of those sets.

	Training data	**Isolated**		**Sentences**	
		Dysarthric	Control	Dysarthric	Control
SGMM	Both	56.1	19.4	41.5	4.4
LF-MMI	Both	**49.2**	24.0	**25.9**	7.9
	Dysarthric	55.0	41.9	42.1	18.4
	Control	52.9	**17.9**	48.7	**2.9**

We observe similar patterns in our experiments on UA-Speech as shown in Figure 4.1. The LF-MMI system overall performs better than the GMM, except for the mildly dysarthric and control speakers. In this case there is not a big performance difference on dysarthric speech between training on only dysarthric or all of the speakers, probably because UA-Speech contains much more data than Torgo. However, on control speech there is still a drop in WER when training on both sets of speakers compared to a control-only system.

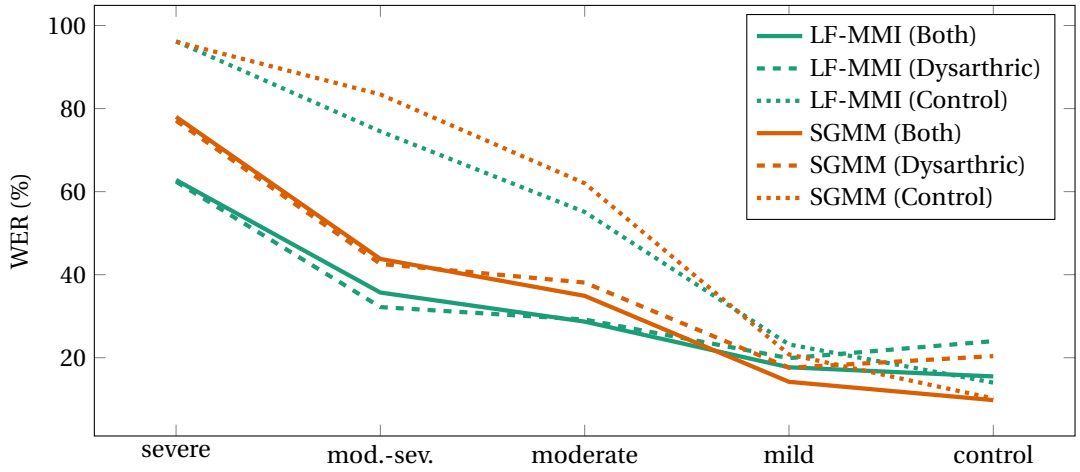


Figure 4.1: WER results on UA-Speech: Training only on dysarthric speakers, only on control speakers, or both sets for SGMM (orange) and LF-MMI (green) systems.

4.4.2 Model combination

Table 4.2 shows the results of model combination of the different LF-MMI systems for Torgo from Table 4.1. We find that the combined systems improve substantially on dysarthric speakers for isolated word recognition and are better than any of the individual results. Combining all three systems is best with a WER of 42.2% compared to the previous best of 49.2% in the system trained on both dysarthric and control speech. There are no improvements for dysarthric speech on the sentence task, but as long as the system trained on both sets of speakers is

included, the results are close to the previous one. As hypothesised, the model combination approach allows to also improve on the control speakers, in one instance even outperforming the previous best result. As long as the model trained only on control speakers is included in the ones to combine, the results are better than from that system alone.

Table 4.2: WER results on Torgo, averaged over dysarthric and control speakers, respectively. Lattice combination of the LF-MMI models from Table 4.1.

Combinations of systems from Table 4.1	**Isolated**		**Sentences**	
	Dysarthric	Control	Dysarthric	Control
Dysarthric + Both	44.0	26.9	27.1	10.3
Control + Both	45.1	**16.6**	27.3	4.3
Dysarthric + Control	48.0	21.3	38.1	5.9
All 3 models	**42.2**	19.1	29.0	7.4

Similarly, Figure 4.2 shows absolute WER improvements over the baseline LF-MMI model trained on both dysarthric and control speakers of UA-Speech from Figure 4.1. We observe that when we include the system trained on only control speakers in the ones to combine, the good performance on those speakers is again maintained. The performance on dysarthric speakers can even improve slightly and we see the best results when combining all three LF-MMI systems from Figure 4.1 (32.7% WER across the dysarthric speakers, 13.8% on control).

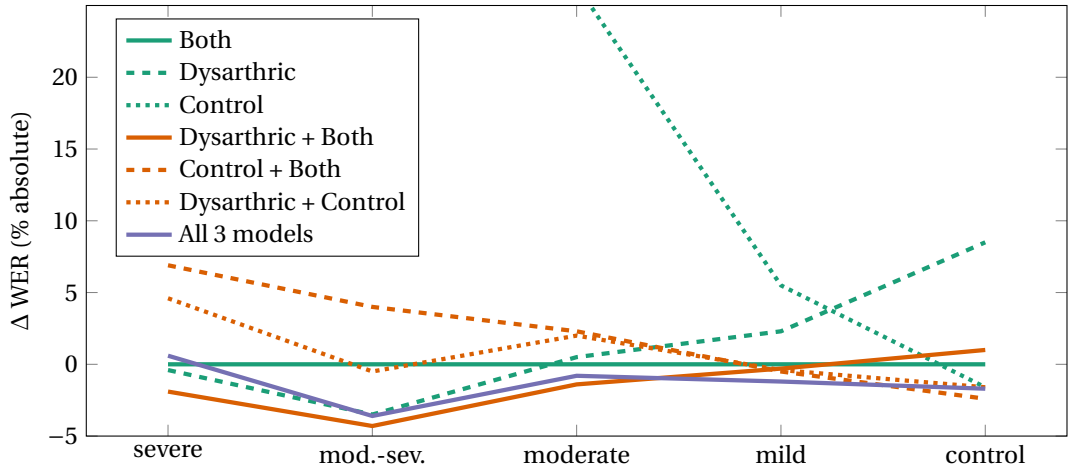


Figure 4.2: WER results on UA-Speech: Lattice combination (orange and purple) of the LF-MMI systems from Figure 4.1, which were trained on only dysarthric speakers, only control speakers, or both sets (green). Shown is the absolute WER difference to the model trained on both dysarthric and control speech (solid green line).

4.4.3 Severity-conditioned models

It appears tempting to take this model combination strategy further and not only combine separate acoustic models trained on dysarthric and control speakers. We also consider training

5 separate severity-specific models on only the data of the respective severity (or the control speech), bridging the gap to fully speaker-dependent systems. We evaluate each of the 5 models on the corresponding test data, which assumes knowing the test speaker's severity in advance. As Table 4.3 shows (*5 separate*), except on the less variable control and mildly dysarthric speech, these individual models perform worse than the models trained on all dysarthric or all data. We hypothesise that this is because the *Dysarthric* and *Both* models benefit both from having more data overall and from similarities between the different severity levels that they can exploit.

Table 4.3: WER results on UA-Speech: We train 5 separate acoustic models for each severity and the control speakers and evaluate them on the corresponding test data. This is better than combining these 5 models. Neither of these beat the LF-MMI systems from Figure 4.1 (green).

Model	Sev.	Mod.-sev.	Mod.	Mild	Control
Baselines					
Both	62.8	35.7	**28.7**	**17.7**	15.5
Dysarthric	**62.4**	**32.2**	29.2	19.9	24.0
Control	96.2	74.5	55.1	23.2	**14.0**
Severity-specific models					
5 separate	70.4	36.8	31.1	18.0	**14.0**
5 combined	78.4	43.6	35.2	19.7	16.2

However, in this case there is also no improvement from combining these 5 models (*5 combined*). This is easily explained because the individual models do not generalise well to other data and yield high error rates on speakers of different severity, which does not leave much room for improvement in the combined model.

4.4.4 Alternative pronunciation models

Dysarthric speech

The previous models were all trained with a phonemic pronunciation lexicon. Figure 4.3 now compares systems for the UA-Speech corpus trained with either a phoneme or grapheme dictionary on both dysarthric and control speech. While phonemes clearly are better suited for the GMM systems, there is only a small gap between phonemes and graphemes with LF-MMI models, with WERs of 33.9% and 35.4%, respectively, on the dysarthric speakers. When we combine these two acoustic models, we obtain our best results so far (31.2% across the dysarthric speakers) and also see a lower WER on control speakers, although still not matching the GMM systems.

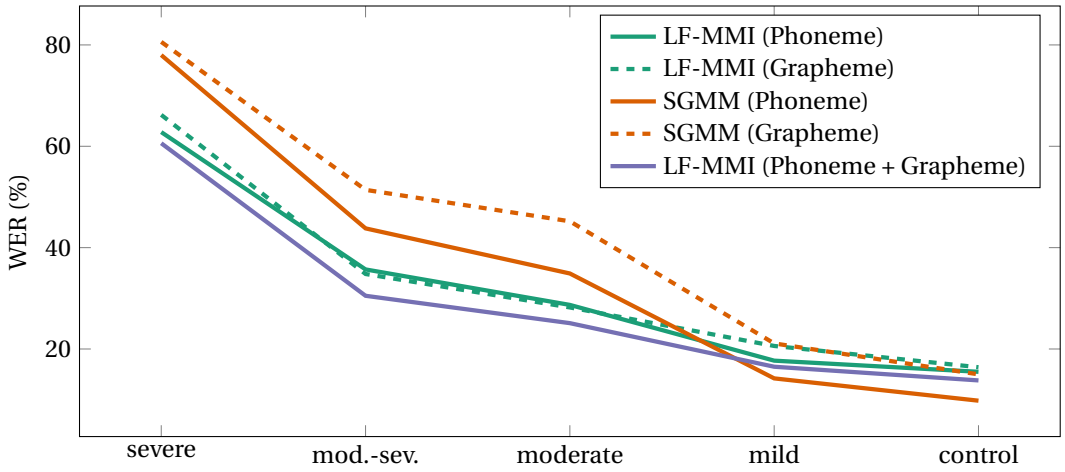


Figure 4.3: WER results on UA-Speech: Comparison of phoneme- and grapheme-based systems and lattice combination of the two. All models are trained on both dysarthric and control speech.

Children's speech

Table 4.4 shows the results of this approach for the PF-STAR children's speech corpus. Our results are significantly better than previous work that is directly comparable (Dubagunta et al., 2019) even though they train on additional adult speech data from the WSJCAM0 corpus (Robinson et al., 1995). When we do the same for our phoneme-based system, our results improve further. We report the rest of our results without external data.

Here, phonemes slightly outperform graphemes by 0.2–0.6% absolute. Model combination of the phoneme- and grapheme-based systems shows only marginal improvements on the test sets. However, it should be noted that the baseline is already very strong — these are among the best results reported on PF-STAR, although direct comparisons are not always possible because of different splits or evaluation protocols.

Table 4.4: WER results for the PF-STAR corpus, comparing phoneme- and grapheme-based systems and lattice combination of the two. For reference, we include the best results of Dubagunta et al. (2019).

Model	dev	test-A	test-B
Dubagunta et al. (2019)		12.0	13.8
Phonemes	**4.9**	6.8	7.2
+ *WSJCAM0*	*4.4*	*6.5*	*6.9*
Graphemes	5.5	7.0	7.5
Combination	5.0	**6.7**	**7.1**

Non-native children's speech

Table 4.5 shows the WERs for the Interspeech 2021 shared task on ASR for non-native children's speech. For English, phonemes clearly perform better than graphemes (by 3.2% absolute). This indicates that despite the non-native speech the relationship between graphemes and phonemes might be too weak in English. For German, which has a more direct mapping from graphemes to phonemes, it is the opposite. Graphemes outperform phonemes by 5.8% absolute. However, model combination of the two does not offer any advantage in either language.

Table 4.5: WER results for the Interspeech 2021 shared task on ASR for non-native children's speech, comparing phoneme- and grapheme-based systems and lattice combination of the two.

Model	**English**		**German**	
	dev	test	dev	test
Phonemes	**13.3**	**33.2**	49.8	45.8
Graphemes	15.5	36.4	**43.7**	**40.0**
Combination	**13.3**	33.3	**43.7**	41.9

4.5 Summary

In this chapter, we aimed to address the acoustic and lexical variability of dysarthric speech with model combination. We found that combining ASR systems trained on different groups of speakers can improve recognition results on dysarthric speech and partially offset the drop in performance on control speech observed when training models on only dysarthric or both dysarthric and control speakers compared to a model trained on control speakers only. We found this to be the case for isolated word recognition on the Torgo and UA-Speech corpora, but on the Torgo sentence task we did not see further improvements on dysarthric speech.

Dysarthric speech, children's speech and non-native speech all have high pronunciation variability. For languages with a shallow orthography like German, grapheme-based ASR systems help to address this, while for English it is not beneficial. Our alternative pronunciation model approach, where phoneme- and grapheme-based systems are combined dynamically during decoding, improves ASR performance in certain cases, especially for dysarthric speech. We also obtain state-of-the-art results on the PF-STAR corpus of children's speech.

Model combination thus provides a good method to handle the acoustic and lexical variations between dysarthric and control speakers and could pave the way for ASR systems that can deal well with a wider range of speech conditions.

5 Discriminability analysis

5.1 Introduction

A successful ASR system needs to discriminate words and word sequences. For this, an acoustic model also has to be able to discriminate individual acoustic units. The decision tree employed in HMM-based ASR approaches aids with this by learning a mapping from phonemes to a discriminable set of clustered, context-dependent phonemes. However, some types of speech data can make this task more challenging. Dysarthric speakers have less control over their articulation and in this chapter we analyse whether this results in a less discriminable acoustic unit space by building upon the findings from the previous chapters.

Razavi and Magimai.-Doss (2015) and Razavi et al. (2018) proposed a method for automatically deriving acoustic subword units and pronunciations from graphemes, for example for low-resource languages where pronunciation dictionaries are not available. To validate that their derived subword units are similar to phonemes, they compared Gaussian distributions estimated for each type of unit with the KL divergence and visualised the resulting confusion matrix. We adopt this analysis approach to compare the acoustic space for the same set of units obtained from different sets of data, such as different groups of speakers.

Rudzicz et al. (2012a) analysed the acoustic and articulatory unit space in a similar manner to assess the similarity of the spaces obtained from different types of features and different subsets of the data. They estimated two Gaussian distributions per phoneme, separately for dysarthric and for control speakers and computed the mean of the KL divergences between each pair of Gaussians. They found that the unit spaces of dysarthric and control speakers were more similar to each other with articulatory features than with acoustic features. Furthermore, they showed that transforming control speaker features to match those of dysarthric speakers increased the similarity of the resulting unit spaces.

We follow the approach of Razavi and Magimai.-Doss (2015) and Razavi et al. (2018) to compare two sets of acoustic units, represented as Gaussian distributions, where we compute the KL divergence between each unit to obtain a confusion matrix. This represents the discriminability

of the acoustic unit space and can be summarised by its median value. Acoustic models are trained to distinguish between acoustic units, so we expect them to perform better when their training data is more discriminable. In this way, we compare the data from only dysarthric speakers, only control speakers, and the combined speaker set. We further extend this to a more realistic task of comparing words, represented by their acoustic unit sequences. We relate the proposed measures to differences in dysarthric speakers' intelligibility and analyse whether they could also be used to compare TTS systems, which are commonly evaluated for intelligibility as well.

We further elaborate on our proposed approach in Section 5.2 and discuss the results in Section 5.3.

5.2 Approach

5.2.1 Comparison of acoustic units

As explained in Section 2.1.1, the set of acoustic units is commonly determined by clustering acoustically similar HMM states with a decision tree. To analyse these units, we then estimate simple Gaussian distributions without mixture components for these units from a given set of training data because the KL divergence between GMMs does not have a closed form solution. The KL divergence $D_{\mathrm{KL}}(f||g)$ between two multivariate Gaussian distributions $\mathcal{N}_f(\boldsymbol{\mu}_f, \Sigma_f)$ and $\mathcal{N}_g(\boldsymbol{\mu}_g, \Sigma_g)$ with mean vectors $\boldsymbol{\mu}$, covariance matrices Σ, and dimensionality d is (Durrieu et al., 2012)

$$D_{\mathrm{KL}}(f||g) = \frac{1}{2}\log\frac{|\Sigma_g|}{|\Sigma_f|} + \frac{1}{2}\mathrm{Tr}(\Sigma_g^{-1}\Sigma_f) + \frac{1}{2}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T\Sigma_g^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g) - \frac{d}{2}\,. \tag{5.1}$$

We compute the KL divergences between all units, resulting in a confusion matrix that we can summarise with the median KL divergence. This allows to compare different sets of acoustic units. We choose the median because the KL divergence is potentially unbounded and we found that large values can sometimes skew the mean. We note that this analysis is only based on the training data and forced alignments, which can be obtained from any basic HMM/GMM acoustic model. It does not require training a dedicated model for the given data and can therefore be carried out before training more resource-intensive neural network models.

5.2.2 Comparison of unit sequences

To better quantify these differences, we demonstrate their effect on a more applied word discrimination task by comparing acoustic unit sequences of word pairs. This closely mirrors

what is expected from ASR systems, especially in the case of sequence-discriminative LF-MMI training, which explicitly learns to discriminate acoustic sequences. We pick a set of words from the training data and find the acoustic unit sequences corresponding to their pronunciation for a given data subset based on the lexicon and decision tree. For each possible pair of two different words, we then compute the dynamic time warping (DTW) distance between the sequences with the KL divergence between the corresponding Gaussians as the local distance, illustrated in Figure 5.1.

The word pair generation and DTW computation is based on code of Kamper (2019).[1]

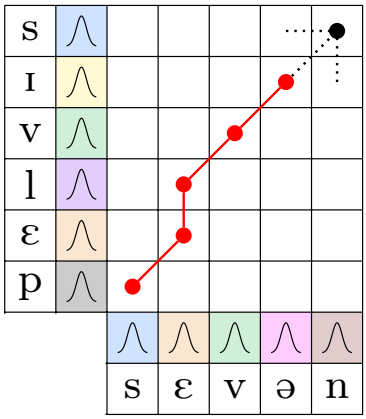


Figure 5.1: Comparison of acoustic unit sequences with dynamic time warping. The local distance is the KL divergence between the Gaussian distributions estimated for the units.

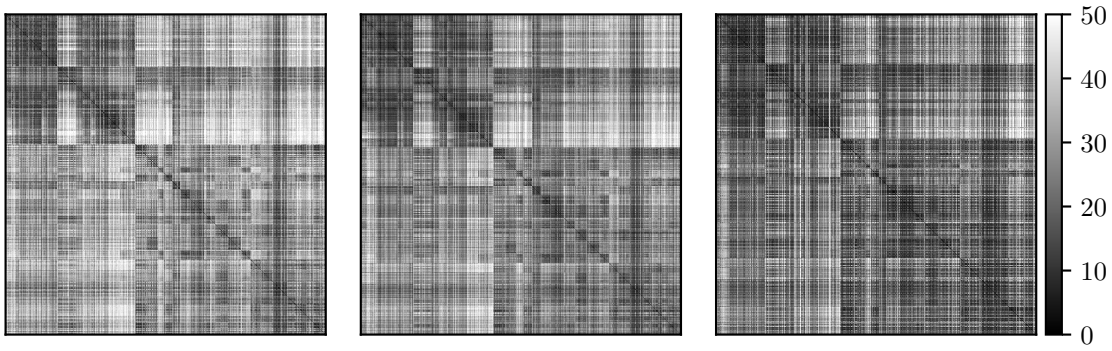
5.3 Results and discussion

5.3.1 Comparison of acoustic units

We first compute the KL divergences between all units for the data from either dysarthric, control, or both sets of speakers from the Torgo corpus that we compared in the previous chapters. In the resulting confusion matrices in Figure 5.2, we observe that the KL divergences, and thus the state and phoneme discriminability, are the highest for control speech. The phonemes are most confusable in dysarthric speech, visible as an increase in darker regions of high similarity, and the combined data falls between the two. Dysarthric speakers are less intelligible and have articulation difficulties, so it is expected that acoustic subword units derived from dysarthric speech are more confusable than those from control speech.

For Figure 5.2, we used the set of acoustic units obtained from decision tree clustering of context-dependent triphones. However, this clustering itself is data-driven and the number of resulting units can vary depending on the amount and the discriminability of the data. For a fair comparison with a consistent number of units, we therefore use monophone units in the

[1] github.com/kamperh/recipe_bucktsong_awe_py3/tree/master/samediff



(a) Control (b) Control + Dysarthric (c) Dysarthric

Figure 5.2: Confusion matrices of clustered context-dependent acoustic units for data from only control, only dysarthric, or both sets of speakers of the Torgo corpus. Units are grouped by manner of articulation and then by phoneme, from consonants in the upper left, to semi-vowels and vowels in the bottom right corner. Darker regions indicate higher similarity and lower KL divergence. Values above 50 are clipped. Including dysarthric speech data results in a less discriminative acoustic unit space where units are more similar to each other.

following analyses, with three Gaussians per phoneme, based on the 3-state HMM topology, resulting in a total number of 122 acoustic units. Table 5.1 shows the median KL divergences obtained in this way from the three data subsets of both Torgo and UA-Speech. It confirms the visual analysis above with the lowest KL divergences, and thus lowest discriminability, on dysarthric speech, highlighting the challenges for dysarthric ASR systems. We observe a bigger difference between dysarthric and control speech on UA-Speech than on Torgo. The effect on the KL divergences of adding speed-perturbed data is negligible. These minor perturbations do not significantly affect the acoustic space, while the additional training data does improve WERs as we showed in Chapter 3.

Table 5.1: Median KL divergences between acoustic units obtained from only dysarthric, only control, or both sets of speakers.

Speakers	**Speed perturbation**	**Torgo**	**UA-Speech**
Dysarthric		12.9	11.9
Dysarthric	✓	12.7	11.7
Control		17.5	23.4
Control	✓	17.3	23.0
Both		14.5	16.3
Both	✓	14.3	16.0

In the case of grapheme units, the median KL divergences are lower than with phonemes, but the overall pattern between different groups of speakers remains, see Table 5.2. This can

be explained by the fact that then there are only 88 total units instead of 122, reducing the overall discriminative power. Nonetheless, we showed in the previous chapter, that WERs of phoneme and grapheme systems trained with LF-MMI are comparable and further improve after model combination.

Table 5.2: Median KL divergences between phoneme and grapheme acoustic units obtained from only dysarthric, only control, or both sets of speakers on UA-Speech.

Speakers	Phonemes	Graphemes
Dysarthric	11.9	10.0
Control	23.4	16.6
Both	16.3	11.8

5.3.2 Comparison of unit sequences

For the word discrimination task, we pick subsets of 289 words from the UA-Speech training data and 1378 words from Torgo with at least four characters and compute the DTW distances between all possible word pairs for monophone acoustic units from the same three subsets of the data as above. Figures 5.3 and 5.5 show histograms of the DTW distances. It is harder to discriminate between words with acoustic units derived from dysarthric speech than those from control speech and combining the data leads to a middle ground. For UA-Speech the differences between speaker groups are more pronounced than for Torgo, mirroring the different ranges of median KL divergences observed in Table 5.1. As described in Section 2.3.1, the Torgo corpus contains many minimal pairs, so it is expected that it is harder in general to discriminate between the words from Torgo than from UA-Speech. In Figures 5.4 and 5.6 we additionally repeat the same word discrimination task with clustered context-dependent instead of monophone units. This allows the decision tree to find a more discriminable set of acoustic units. In this case, the differences between dysarthric and control speech are also clearly apparent on the data from Torgo. This confirms that our previous analysis on discriminability of acoustic units also translates to applied settings like this work discrimination task.

Figure 5.7 shows that also according to the word discrimination task speed perturbation has no significant effect on the acoustic space.

Tables 4.1 and 3.4 showed that while including control speech into LF-MMI training is crucial to perform well on dysarthric speakers, we observed the lowest WERs on control speech when training without any dysarthric speech data. These analyses explain that this is due to the dysarthric speech adding too much variability for the neural network training. This affects the acoustic unit space and reduces the model's discriminative power. We were able to offset this in the previous chapter with model combination. Similarly, we showed in Chapter 3 that data augmentation by speed perturbation — which we use in all our LF-MMI experiments — helps for dysarthric speakers, but not introducing this additional source of variability is

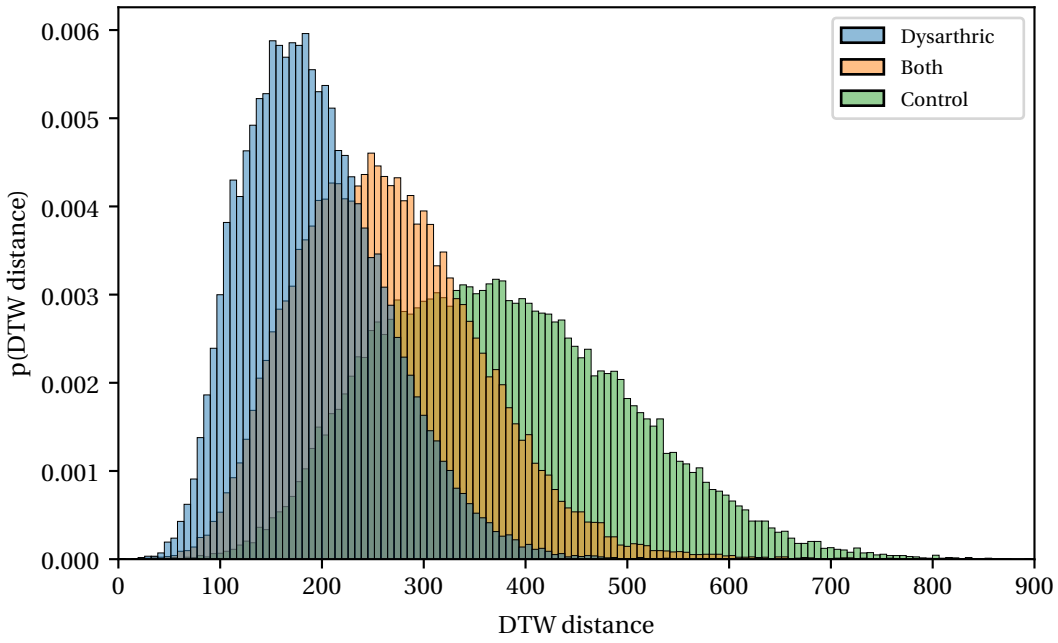


Figure 5.3: Histograms of DTW distances between word pairs with monophone units from only dysarthric, only control, or both sets of speakers of UA-Speech.

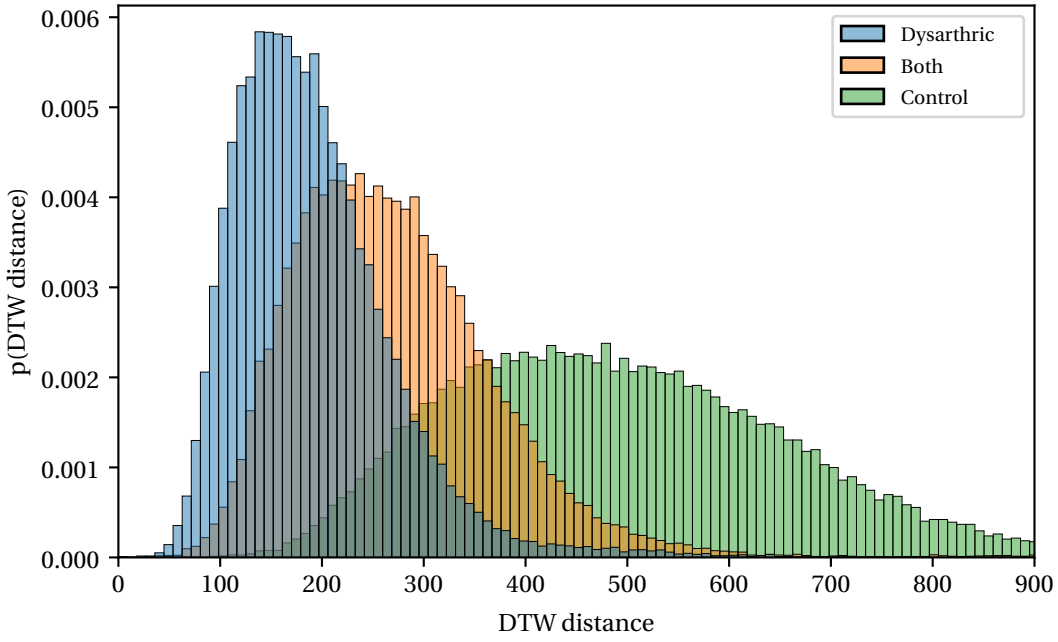


Figure 5.4: Histograms of DTW distances between word pairs with clustered context-dependent units from only dysarthric, only control, or both sets of speakers of UA-Speech.

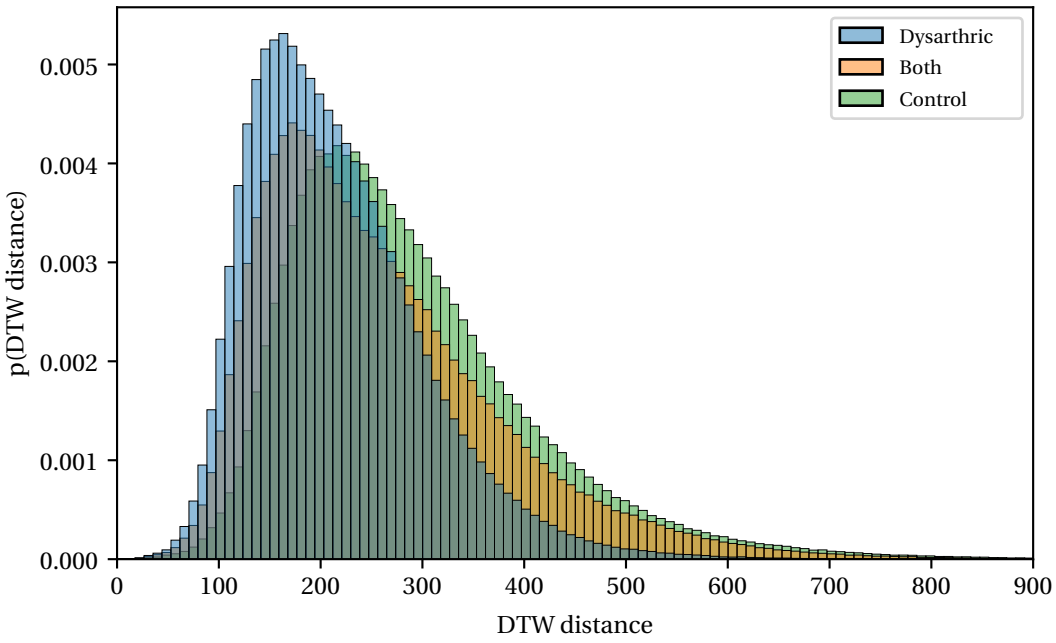


Figure 5.5: Histograms of DTW distances between word pairs with monophone units from only dysarthric, only control, or both sets of speakers of Torgo.

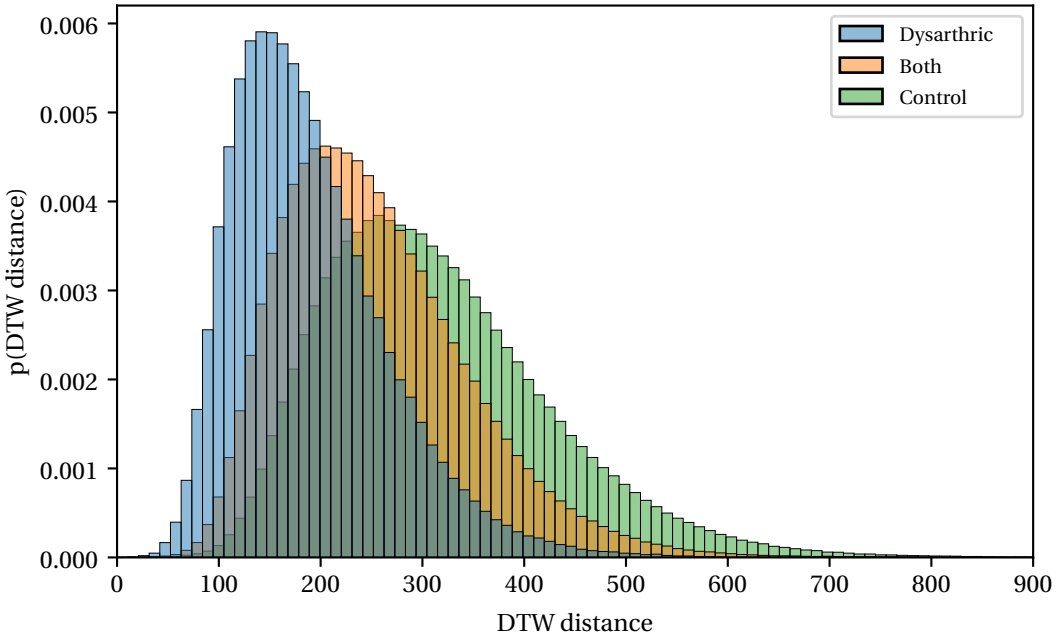


Figure 5.6: Histograms of DTW distances between word pairs with clustered context-dependent units from only dysarthric, only control, or both sets of speakers of Torgo.

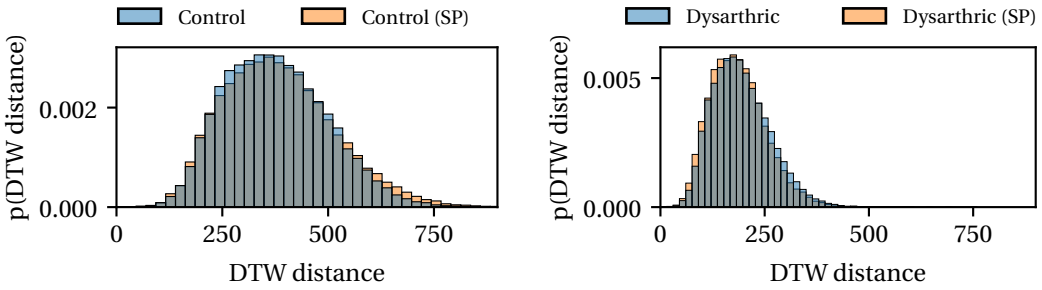


Figure 5.7: Histograms of DTW distances between word pairs with acoustic units from dysarthric and control speech from UA-Speech with and without speed perturbation (SP).

better for control speakers when training on both sets. Our observations are also consistent with the literature, where reduction of vowel space area or effects on related metrics have been observed in dysarthric speech (Turner et al., 1995; Lansford and Liss, 2014). Instead of focusing only on the vowels, we cover all phonemes with our analysis approach.

5.3.3 Relationship to speaker severity and intelligibility

We showed that the acoustic unit space of control speech is more discriminable than that of dysarthric speech. We now examine whether our analysis also picks up on the more fine-grained differences in intelligibility between different dysarthric speakers. We further review whether this approach could also be applied to the comparison of TTS systems for typical speech, which are also commonly evaluated for intelligibility.

Dysarthric speech

We estimate Gaussian distributions for all monophone acoustic units separately for each dysarthric speaker from UA-Speech. We observe a strong correlation (Pearson's r = 0.90) between the resulting median KL divergences and the human intelligibility ratings provided in the corpus, see Figure 5.8. Similarly, higher acoustic discriminability is strongly correlated with lower WERs (Pearson's r = -0.89) in the SD LF-MMI acoustic models that we trained in Chapter 3 on the corresponding speaker's data. For reference, we also show the overall results for control speakers, which fall in the same region as mildly dysarthric speakers in terms of discriminability and ASR results for the LF-MMI model trained only on control speech.

When we additionally add all control speech to each dysarthric speaker's data, the KL divergences increase across the board, in particular for the more severely dysarthric speakers, as shown in Figure 5.9. The ASR performance of the SD + Control systems from Chapter 3 improves accordingly. While the differences in discriminability between speakers are now much smaller, there is still a strong correlation with the human-rated intelligibility (Pearson's r = 0.82).

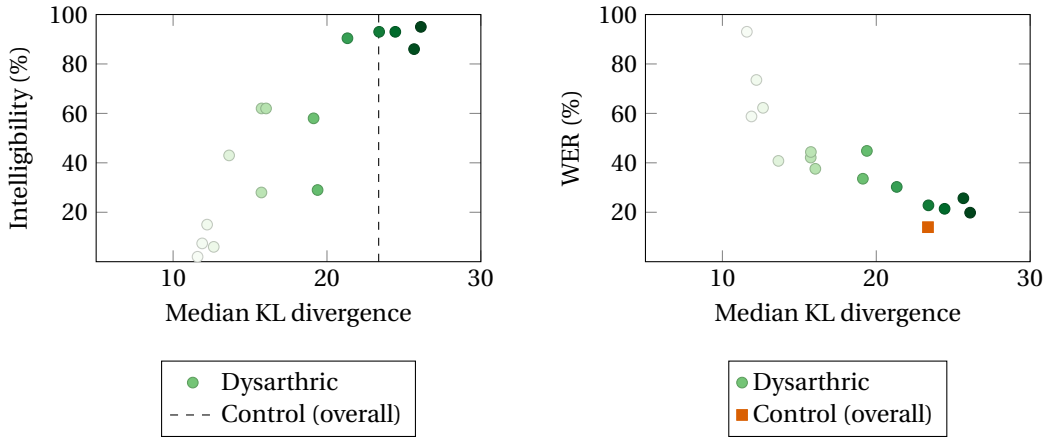


Figure 5.8: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings and WER for each dysarthric speaker from UA-Speech. Higher median KL divergence is strongly correlated with higher intelligibility (Pearson's r = 0.90) and lower WER (r = -0.89).

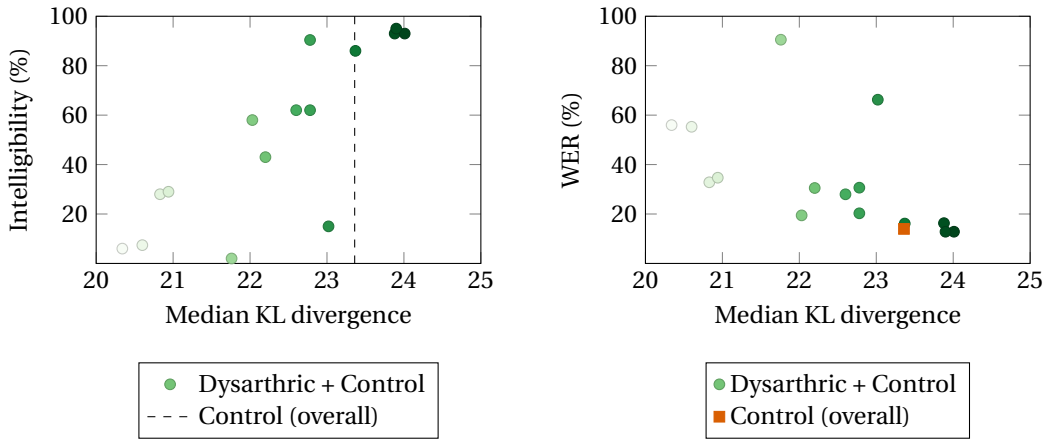


Figure 5.9: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings and WER for each dysarthric speaker from UA-Speech with added control speech. Higher median KL divergence is strongly correlated with higher intelligibility (Pearson's r = 0.82) and lower WER (r = -0.88).

The intelligibility of dysarthric speakers from Torgo was assessed according to the Frenchay Dysarthria Assessment (Enderby, 1980) by a speech and language pathologist. This evaluation was done for words, sentences and in conversation, assigning scores on a 9-point scale. We take the average of these three scores and convert it to percent for consistency with the UA-Speech intelligibility ratings. We then compute the median KL divergence for each speaker in the same way as for UA-Speech above. Figure 5.10 shows their relationship to the human intelligibility ratings (Pearson's r = 0.37) and ASR results on the isolated work task (r = -0.36) of the SI LF-MMI models from Chapter 3. In this case, these measures are only lightly correlated, but we note that there are only eight dysarthric speakers in Torgo as opposed to 15 in UA-Speech.

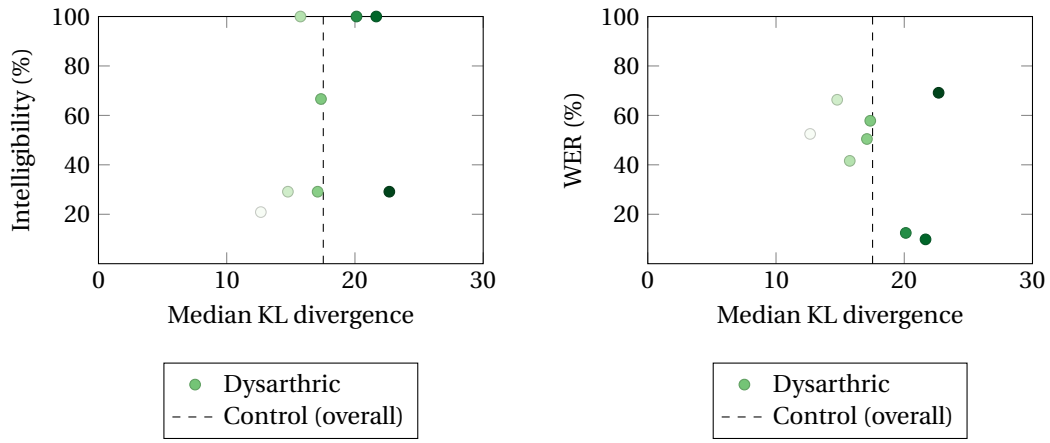


Figure 5.10: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings and WER for each dysarthric speaker from Torgo. The WERs are from speaker-independent LF-MMI models on the isolated word task. Higher median KL divergence is lightly correlated with higher intelligibility (Pearson's r = 0.37) and lower WER (r = -0.36).

Synthetic speech

Can this approach be used to not only analyse the intelligibility of dysarthric speech, but also that of synthesised speech? The Blizzard Challenge is an annual event where TTS systems are evaluated on a shared task. We ran our analysis on the synthetic speech outputs of the 16 systems submitted to the Blizzard Challenge 2016 (King and Karaiskos, 2016). These are based on a variety of TTS methods, including concatenative, statistical HMM, and hybrid DNN synthesis, but do not include the most recent fully neural architectures that can produce very high quality speech. The training data for all systems was 5 hours of English speech from audiobooks for children read in an expressive style. A fixed list of outputs had to be generated, which were then evaluated by human listeners on a range of measures, including naturalness and intelligibility. The latter was evaluated by computing the WER of the listeners' transcriptions of semantically unpredictable sentences (SUS) (Benoît et al., 1996). For each submission, we estimate the Gaussian distributions for monophone units from all the synthesised speech, not only the outputs generated for the SUS task, to ensure there is sufficient data. We also include the provided human recordings of the target outputs as a reference point, which do not have any intelligibility rating. Based on our previous experiments on dysarthric speech, we would expect systems with a lower human-annotated WER, i.e. higher intelligibility, to have higher KL divergences.

Figure 5.11 shows the relationship of median KL divergence between acoustic units and WER on the SUS task. The Pearson correlation coefficient is -0.30, indicating only a low negative correlation, compared to the higher correlations observed on dysarthric speech. We assume this is because the reasons why a TTS system is not intelligible can be manifold and depend on the nature of the system. The Blizzard Challenge submissions are based on many different methods that result in a variety of output characteristics. On the other

hand, reductions in intelligibility in dysarthric speech are more clearly linked to the dysarthria severity and associated articulation difficulties, which is better captured by the acoustic space discriminability.

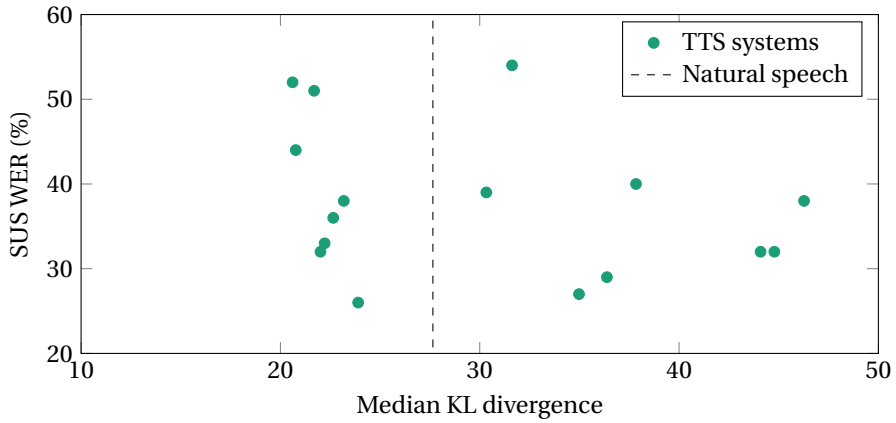


Figure 5.11: Median KL divergence between acoustic units and WER on the semantically unpredictable sentence (SUS) task of 16 TTS systems from the Blizzard Challenge 2016. Higher median KL divergence is only lightly correlated with a decrease in WER (Pearson's r = -0.30).

5.3.4 Analysis of children's speech

We have demonstrated the suitability of our analysis framework for dysarthric speech and the relationship between acoustic discriminability, speaker intelligibility, and ASR performance in the previous sections. It could potentially also be applied to other kinds of atypical speech where the acoustic unit space is affected. For example, children gain phonetic ability and articulatory skills over time (Dodd et al., 2003). Therefore, we hypothesise that the acoustic unit space of younger children is less discriminable than that of older ones.

We take the data of the 80 children from the training set of the PF-STAR corpus and estimate Gaussian distributions for all monophone acoustic units separately for each child. Figure 5.12 shows the resulting median KL divergence for each child and its age. KL divergences are spread widely for children aged 8–10 years. Children younger than 7 have slightly reduced and children older than 11 slightly increased acoustic discriminability, but the dataset contains only few children in these age groups.

There is overall no clear correlation between the KL divergences and age, with a Pearson correlation coefficient of -0.05. This is likely because most children in the dataset are at least 8 years old, when their articulation is already well developed (Dodd et al., 2003). This analysis should therefore be repeated with a dataset of younger children.

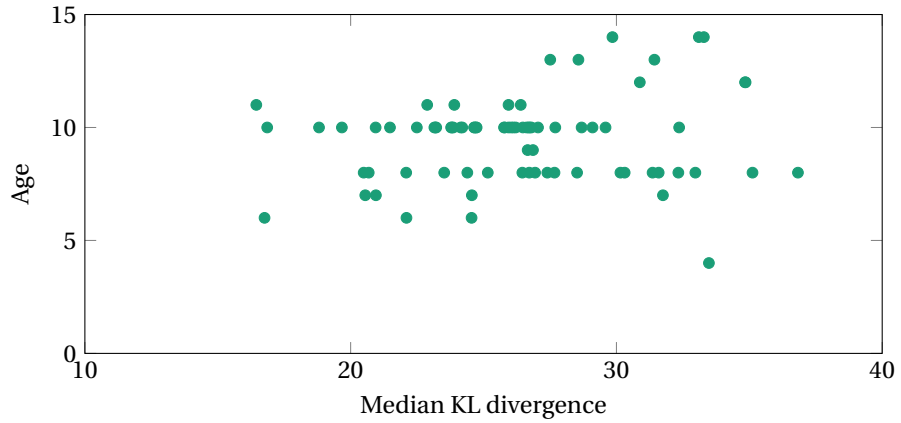


Figure 5.12: Median KL divergence between acoustic units and age of 80 children from the PF-STAR corpus. There is no correlation between the median KL divergences and children's age (Pearson's r = -0.05, Spearman's ρ = 0.05).

5.4 Summary

In this chapter we analysed the acoustic discriminability of different kinds of speech based on KL divergences between acoustic subword units. We found that, as expected, typical speech is much more discriminable than dysarthric speech. Furthermore, our experiments showed that our proposed analysis approach is a good predictor of dysarthric speech intelligibility and of ASR performance and can explain difference between individual dysarthric speakers. However, it does not necessarily extend to intelligibility analysis in other domains, such as TTS, where the reasons for reduced intelligibility are different than for dysarthric speech. More research is required to investigate whether the proposed discriminability measures are also related to other factors, such as age or foreign accent.

6 Data augmentation for dysarthric speech recognition

In the previous chapters, we investigated different methods to better adapt the acoustic model for dysarthric ASR. However, mismatches between dysarthric and typical speech can occur at multiple levels. While some of these can be addressed at the model level, others should be considered at the data level. Recording large amounts of data can also be exhausting for speakers with dysarthria. Model adaptation is therefore not always sufficient because of the inherent scarcity of dysarthric training data. As discussed in Section 2.2.3, data augmentation is an additional option. In Chapter 3, we have already demonstrated speed perturbation as an effective data augmentation method, where we simply add additional copies of the training data that are speed-perturbed by a small factor. In this chapter, we now consider more sophisticated voice conversion (VC) and text-to-speech (TTS) approaches that directly model the specific characteristics of a target speaker and of dysarthric speech.

In Section 6.1, we start with an interpretable signal processing VC baseline that transforms general attributes related to the speaker identity, such as formants, but ignores paralinguistic aspects like pathological conditions. We then evaluate GAN-based VC in Section 6.2, which allows to learn a wider range of speaker characteristics and specifically model dysarthric speech. Finally, in Section 6.3 we consider dysarthric speech synthesis, which does not require source data at synthesis time, so that speech for a wider range of domains can be generated. We additionally evaluate the use of dysarthric TTS in a few-shot learning scenario, where an acoustic model is trained with only very little data from a target speaker. We further analyse the generated speech and corresponding acoustic subword units from these three approaches with the framework developed in the previous chapter to verify if the generated outputs have similar characteristics as the original dysarthric speech.

6.1 Dysarthria-agnostic voice conversion by pseudonymisation

We begin with an interpretable VC baseline based on signal processing (Dubagunta et al., 2020, 2022).[1] It computes a range of statistics for each speaker, including formant values,

[1] https://github.com/robvanson/PseudonymizeSpeech

fundamental frequency and speaking rate, and then transforms the source speech to match the target speaker's statistics. It does not take into account paralinguistic aspects and focuses on changing the speaker identity, so pathological conditions are not affected by VC. In the previous chapters we learned that including control speech when training dysarthric ASR systems helps to improve their acoustic discriminability and performance. Several studies (Vachhani et al., 2018; Xiong et al., 2019; Geng et al., 2020) have also found minor benefits from additionally adjusting the speaking rates of control speakers to match those of the target dysarthric speaker. In this section, we evaluate whether it is further beneficial to also transform the source speakers' identity to match that of a target dysarthric speaker.

6.1.1 Background

Dubagunta et al. (2020, 2022) developed this VC method for the VoicePrivacy 2020 Challenge (Tomashenko et al., 2022). Its aim is pseudonymisation, i.e. a reversible anonymisation of the source speech for privacy-sensitive applications. For example, the voice of a patient with dysarthria could be recorded as part of a research study, but would then only be shared publicly in pseudonymised form. If necessary, the original speech could be recovered by authorised persons with a hidden set of parameters.

The pseudonymisation pipeline is illustrated in Figure 6.1. A set of speaker characteristics, including formant values, fundamental frequency, and speaking rate, is estimated from the speech of source speaker A. The pseudonymiser then transforms the speech of speaker A based on their characteristics and those of another target speaker B to sound like speech from speaker B. This process can be repeated in reverse to recover speech sounding like that of speaker A by someone with access to that speaker's parameters.

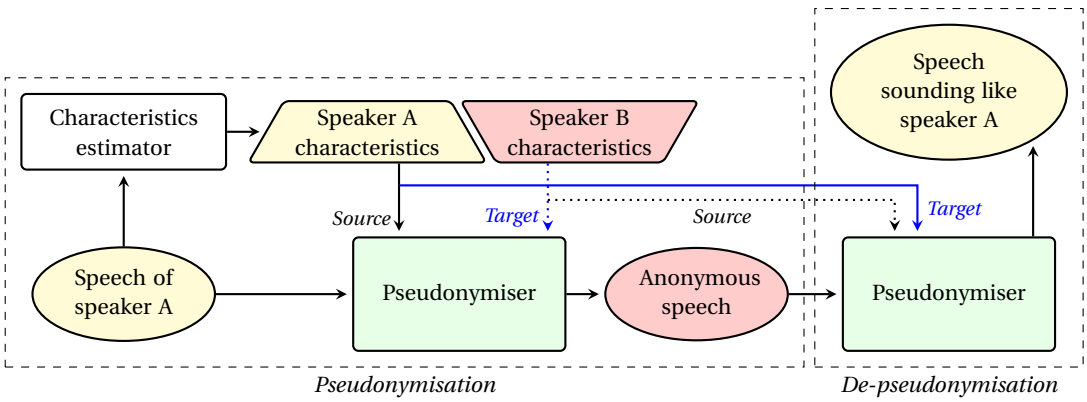


Figure 6.1: Overview of signal processing-based speech pseudonymisation. Figure adapted from Dubagunta et al. (2022).

The pseudonymisation procedure is implemented in Praat (Boersma and Weenink, 2021). It changes the speaking rate and fundamental frequency of the speech and then simulates a different vocal tract length based on the source and target speaker's characteristics with signal

processing algorithms. The pseudonymiser does not require training and all its parameters are interpretable. The output can also be manipulated by manually setting the target speaker characteristics. For further details, we refer the reader to Dubagunta et al. (2022).

We note that one of the explicit aims of this pseudonymisation method is to preserve paralinguistic aspects of the source speech, so that dysarthria characteristics of the source speaker should not be affected when adapting the voice to an unimpaired target speaker. Conversely, when transforming from a control to a dysarthric speaker's voice, we would not expect to introduce dysarthria characteristics, other than changes to the speaking rate. Dubagunta et al. (2022) confirmed this by evaluating the pseudonymised output with a dysarthria classification model.

6.1.2 Experimental setup

In this section, we use SD LF-MMI acoustic models trained on UA-Speech as described in Chapter 3. We compare the SD models on their own, with added control speech, with added tempo-matched control speech, and with added pseudonymised control speech. Furthermore, we compare with dysarthric speech from the other speakers as source speech for pseudonymisation instead of control speech. Speed perturbation with factors $\{0.9, 1.0, 1.1\}$ is additionally used for all models.

To obtain tempo-matched control speech, we compute the mean duration T_i of non-silence phonemes for each speaker i from forced alignment of the training data with the SD HMM/GMM ASR system, following Xiong et al. (2019). Across all control speakers, the mean duration is 135 ms, compared to 209 ms for dysarthric speakers. Figure 6.2 additionally shows the association between mean phoneme duration and the subjective intelligibility ratings of each dysarthric speaker. For each dysarthric speaker d, we then calculate the speech tempo ratio

$$R_{C \to d} = \frac{T_C}{T_d} \tag{6.1}$$

between the mean duration T_C of all control speakers and the mean of speaker d. With these ratios and the `sox` command line tool, we can adjust the tempo of the control speaker utterances to match that of a target dysarthric speaker, while preserving the original pitch and spectral characteristics.

We compute the required speaker statistics for pseudonymisation from each UA-Speech speaker with the provided Praat script. For each dysarthric speaker, we then convert the data with the pseudonymiser from all other dysarthric and control speakers towards this target speaker. This speaker-specific pseudonymised data can be added when training SD acoustic models. We do not modify the default settings of the pseudonymiser. Additionally, we analyse the acoustic units of the pseudonymised data with the framework developed in Chapter 5.

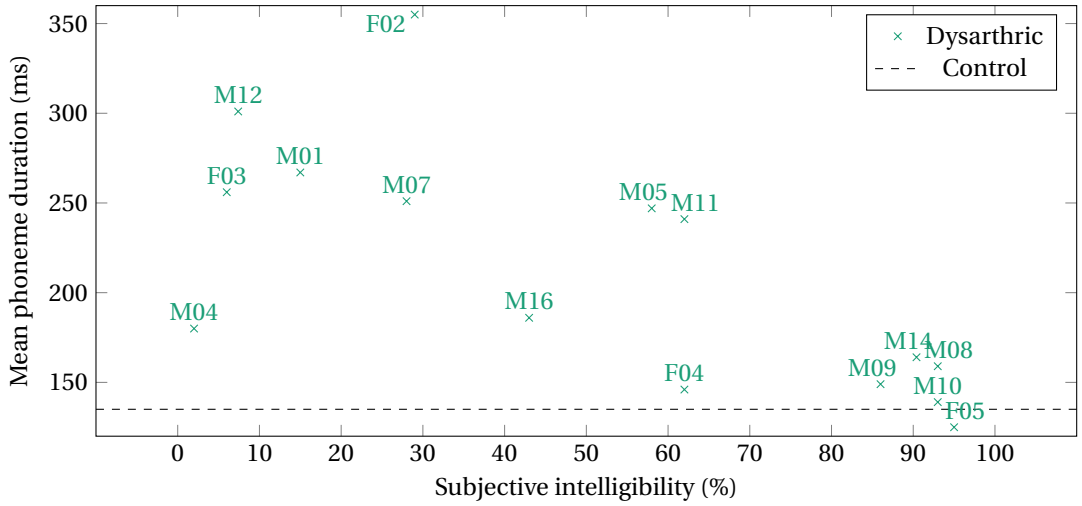


Figure 6.2: Mean phoneme durations and subjective intelligibility ratings for each UA-Speech dysarthric speaker. The dashed black line shows the mean phoneme duration across all control speakers.

6.1.3 Results and analysis

The ASR results are shown in Table 6.1. We first include three baselines from Chapter 3: SD acoustic models trained only on speech from the target dysarthric speaker, SD models additionally trained on all control speech, and a single model trained on all dysarthric speech.

Table 6.1: WERs of LF-MMI acoustic models for each group of dysarthric speakers of UA-Speech when augmenting the training data with dysarthria-agnostic voice conversion (pseudonymisation).

Systems	**Severe**	**Mod.-Severe**	**Moderate**	**Mild**	**Overall**
Baselines from Chapter 3					
SD	70.3	42.7	38.2	24.0	41.3
SD + Control	65.5	32.8	**25.8**	**15.7**	**32.6**
Dysarthric	**62.4**	**32.2**	29.2	19.0	34.0
Data augmentation					
SD + Tempo-matched Control	70.3	33.0	28.2	17.2	34.8
SD + Pseudo-Control	68.6	32.9	30.7	17.7	35.1
SD + Pseudo-Dysarthric	65.9	34.8	37.5	23.2	38.2

When adjusting all control speech to match the target dysarthric speaker's speech tempo (SD + Tempo-matched Control), we observe slightly worse results than adding the original control speech although it is still better than the SD model alone. This is different from Xiong et al. (2019), who observed a small absolute WER improvement of 0.6%. Adding pseudonymised control speech (SD + Pseudo-Control), which not only matches the speaking rates, but also the speaker identity, performs similar to tempo-matched control speech. Pseudonymised

dysarthric speech performs even worse than pseudonymised control speech when added to the SD models (SD + Pseudo-Dysarthric). It is also worse than just adding the remaining dysarthric speech without any transformation (Dysarthric).

We compute the median KL divergences between monophone acoustic units as described in Chapter 5 for speech pseudonymised from either control or dysarthric speech to each dysarthric speaker. Figure 6.3 shows the relationship between these median KL divergences and the subjective intelligibility rating of each speaker. As expected, converting from control speakers results in more discriminable speech with higher KL divergences. There are no major differences in discriminability between each target speaker within either group, confirming that pseudonymisation does not affect paralinguistic aspects related to the pathological condition.

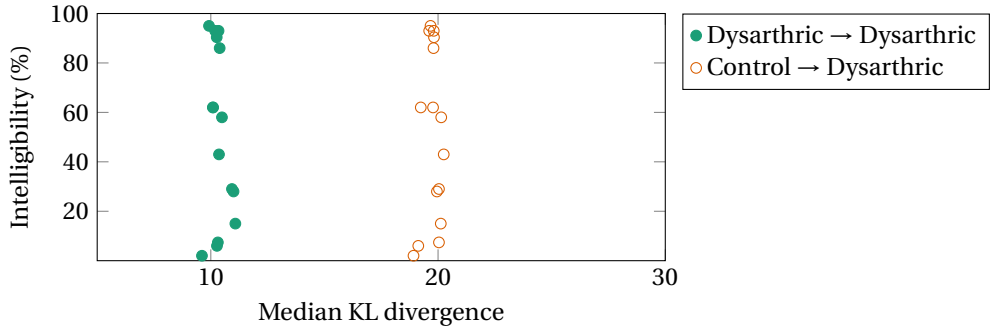


Figure 6.3: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings of control and dysarthric speech from UA-Speech that was converted to other dysarthric speakers via pseudonymisation.

6.2 GAN-based voice conversion

Neural network-based VC is another increasingly popular data augmentation method for dysarthric speech recognition. A mapping from unimpaired control to dysarthric speakers or between different dysarthric speakers is learned, so that additional speech for ASR training can be generated. This requires that recordings of these target utterances are available as a source. However, depending on the type of VC system, parallel training data is not always necessary. Non-parallel methods can be trained without source and target speaker recording the same utterances. Existing applications to dysarthric ASR have been restricted to VC models that convert only between single pairs of speakers, although in general many-to-many VC approaches also exist (Kaneko et al., 2019b).

We note that VC can also be used as a form of feature adaptation, where the voice of dysarthric speakers is converted to that of control speakers, so that it could be fed into an existing ASR system trained only on control speech. Prananta et al. (2022) conclude that in this case most of the benefits from GAN-based VC can already be achieved by simple time stretching of the dysarthric speech to match control speakers' speaking rates. We do not consider this in this

thesis and only convert from control to dysarthric speech. Another approach could be to convert between different dysarthric speakers (Illa et al., 2021), which has not been used for data augmentation in dysarthric ASR yet.

6.2.1 Background

For our experiments, we choose the MaskCycleGAN-VC (Kaneko et al., 2021) architecture[2] that is based on CycleGAN-VC (Kaneko and Kameoka, 2018) and its extension CycleGAN-VC2 (Kaneko et al., 2019a), based on prior comparisons of the output quality of these systems (Prananta et al., 2022). All of these architectures are similar and trained with an adversarial (Goodfellow et al., 2014) and a cycle-consistency loss (Zhu et al., 2017) on non-parallel data, illustrated in Figure 6.4.

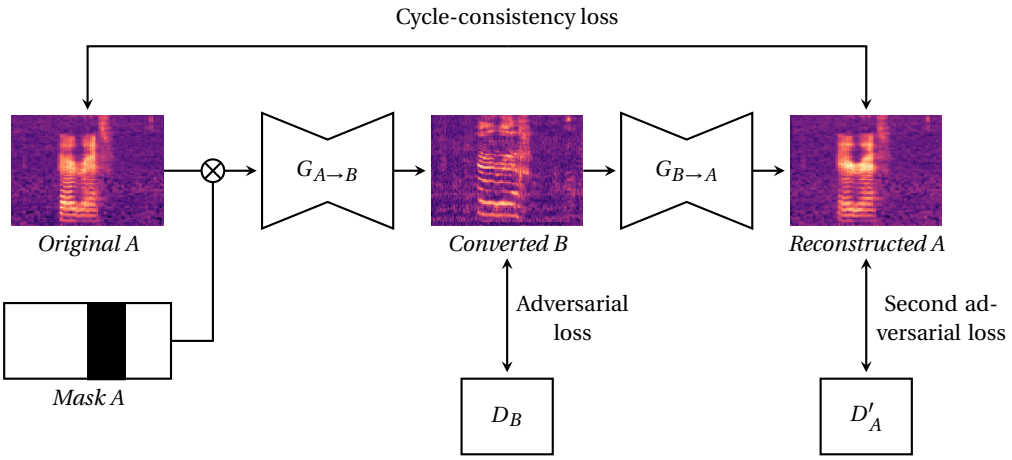


Figure 6.4: Overview of MaskCycleGAN voice conversion training between speakers A and B with generators $G_{A\to B}$ and $G_{B\to A}$ and discriminators D_B and D'_A. Figure adapted from Kaneko et al. (2021).

More precisely, MaskCycleGAN-VC consists of two generator and two discriminator models. The generator $G_{A\to B}$ converts the speech from speaker A to sound like that of speaker B and the second generator $G_{B\to A}$ reconstructs the original speech signal of speaker A based on the converted sample. The discriminator D_B then tries to distinguish converted samples from real samples of speaker B, while the second discriminator D'_A tries to distinguish the reconstructed from real samples of speaker A. Each discriminator is trained to minimise their respective adversarial loss while the generators are trained to increase it and produce more convincing outputs. Because there is no parallel training data, the discriminators focus only on the speaker identity. The cycle-consistency loss is added in order to still preserve the linguistic content – it ensures that the reconstructed signal is as close as possible to the original one. MaskCycleGAN-VC additionally masks a portion of the input to $G_{A\to B}$ along the time axis, which the generator is then forced to fill in, as a form of self-supervised learning for improved

[2] https://github.com/GANtastic3/MaskCycleGAN-VC

modelling of time-frequency structure in the output (Kaneko et al., 2021). For further details on the model architecture and training procedure, we refer the reader to Kaneko et al. (2021).

6.2.2 Experimental setup

We continue to use UA-Speech for these data augmentation experiments. In line with prior work (Prananta et al., 2022), we only convert from male control to male dysarthric and from female control to female dysarthric speakers to simplify the task. This means that there is overall more additional training data for male dysarthric speakers due to the gender imbalance of UA-Speech.

Our VC method only takes into account the changes in speech characteristics between speakers. It does not adjust the duration of an utterance, although we know that the speaking rates of dysarthric speakers can be significantly lower than those of typical speakers. Therefore, we add a separate time stretching step before VC training, similar to previous work (Halpern et al., 2021). We compute speech tempo ratios based on mean phoneme durations for each dysarthric speaker and averaged over all control speakers as described in Section 6.1.2. We then adjust the speech tempo of control speech specifically for each dysarthric speaker based on these ratios with the `sox` command line tool.

We train a separate MaskCycleGAN-VC model for each male-male and female-female pair of control and dysarthric speakers from UA-Speech on all data of block 1 for 100 epochs with a batch size of 1. The input and output features are normalised 80-dimensional Mel spectrograms. We generate waveforms from the voice-converted outputs for both blocks 1 and 3 with a pre-trained MelGAN vocoder (Kumar and Kumar, 2019) model[3] at a sampling rate of 22050 Hz that we downsample to 16 kHz for ASR training.

For ASR, we again train LF-MMI acoustic models on UA-Speech as described in Chapter 3. We compare the previous baselines with adding voice-converted data to them. Speed perturbation with factors $\{0.9, 1.0, 1.1\}$ is applied in all models. Finally, we analyse the acoustic units of the voice-converted data with the framework developed in Chapter 5 to determine whether dysarthric speech characteristics are reproduced.

6.2.3 Results and analysis

We first train SD acoustic models only on the data that was voice-converted from control to dysarthric speakers (Table 6.2, *VC only*). These lead to lower WERs on more severely dysarthric speakers compared to a model trained only on the original control speech (*Control*), while WERs for mild to moderate dysarthric speech increase.

We then train SD models where we augment the target speaker's speech with the VC data converted from all same-gender control speakers to the target speaker (*SD + VC*). This performs

[3] https://github.com/descriptinc/melgan-neurips

Table 6.2: Word error rates (WER) for each group of dysarthric speakers from UA-Speech for data augmentation with MaskCycleGAN-VC.

Systems	Severe	Mod.-Severe	Moderate	Mild	Overall
Baselines from Chapter 3					
SD	70.3	42.7	38.2	24.0	41.3
SD + Control	65.5	32.8	**25.8**	**15.7**	**32.6**
Dysarthric	**62.4**	**32.2**	29.2	19.0	34.0
Control	96.2	74.5	55.1	23.2	56.9
Both	62.8	35.7	28.7	17.7	33.9
Data augmentation					
SD + Tempo-matched Control	70.3	33.0	28.2	17.2	34.8
VC only	90.8	67.9	57.6	32.3	58.1
SD + VC	69.2	37.9	36.5	25.3	40.2
SD + Control + VC	65.4	35.0	29.9	20.5	35.6
Dysarthric + VC	74.0	45.0	40.9	30.8	45.6
Both + VC	74.5	45.6	38.8	27.9	44.4
Both-MelGAN	74.9	37.8	36.0	22.8	40.4

considerably worse than just adding all original control speech (*SD + Control*), with an overall WER of 40.2% compared to 32.6% across all dysarthric speakers. Additionally adding the original control speech (*SD + Control + VC*) also still performs worse at 35.6%. Next, we combine all original dysarthric and all converted data (*Dysarthric + VC*). Finally, we also add the original control speech to this (*Both + VC*). In both cases, adding the VC data leads to worse results than the baseline.

To understand the source of these poor results, we measure the effect of the MelGAN vocoder in the VC experiments. We convert both the control and dysarthric UA-Speech training data to Mel spectrograms and feed these through the MelGAN vocoder. Then we train a model on this data (*Both-MelGAN*) and compare it to the one trained on the original speech (*Both*). Across all dysarthric speakers, we observe an absolute WER increase of 6.5% that we can also expect when training on voice-converted data.

The median KL divergences between monophone acoustic units of the converted control speech are strongly correlated with the subjective intelligibility ratings for the respective target dysarthric speaker (Pearson's $r = 0.92$), as can be seen in Figure 6.5. This highlights that the VC models learn meaningful mappings from control to dysarthric speech and reproduce the differences in acoustic space discriminability between speakers that we observed in the original dysarthric speech ($r = 0.90$). The median KL divergences of tempo-matched control speech, which is the source speech for VC, are also correlated with the subjective intelligibility ratings ($r = 0.53$), but they are still clustered close around the value of the original control speech, indicating no significant reduction in discriminability.

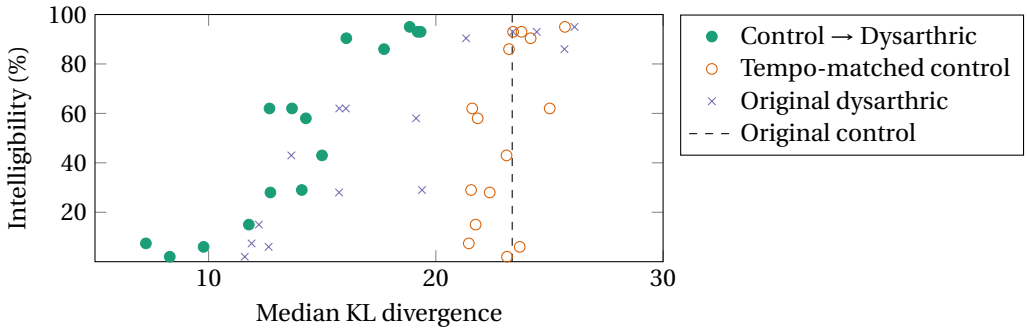


Figure 6.5: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings of control speech from UA-Speech that was converted to dysarthric speakers with MaskCycleGAN-VC (Pearson's $r = 0.92$), of tempo-matched control speech ($r = 0.53$), and of the original dysarthric speech ($r = 0.90$). The dashed line shows the median KL divergence across all the original control speech.

6.3 Speech synthesis

In this section we investigate augmentation of the training data with a TTS system. In contrast to VC, which requires existing recordings of the target utterances, we can synthesise speech for arbitrary sentences and therefore quickly adapt an ASR system to new commands and domains. TTS-based data augmentation has already been applied to ASR for low-resource languages and children's speech (Kadyan et al., 2021). We go beyond using TTS only for data augmentation, but also explore it in a few-shot scenario where we synthesise dysarthric speech based on only a few recordings of a target speaker that was not seen during TTS training.

Few-shot and even zero-shot approaches to pathological speech recognition can be successful (Xiao et al., 2022; Green et al., 2021; Tobin and Tomanek, 2022). Out of the box, a very large acoustic model with up to 10 billion parameters trained on 4.5 million hours of speech (Xiao et al., 2022) reaches state-of-the-art performance on AphasiaBank (MacWhinney et al., 2011), a database of aphasic speech. Fine-tuning on this data gives a further 50% relative improvement. However, such amounts of training data are only available to a few private companies. Even fine-tuning and applying a pretrained model with so many parameters is challenging and storing personalised models for each speaker is costly (Tomanek et al., 2021). It is therefore desirable to also investigate more moderately sized models and alternative few-shot approaches.

In this section we build upon previous work on TTS for dysarthric speech (Soleymanpour et al., 2022) based on the FastSpeech 2 TTS system (Ren et al., 2021). It is a multi-speaker TTS model that is trained on many speakers and is thus better able to capture their diversity. It also simplifies the training procedure with respect to VC-based data augmentation where we trained over one hundred separate models to cover all speaker pairs. A further advantage is that FastSpeech 2 directly learns to model the speech duration and a separate time stretching step to adjust control speech to dysarthric speaker's speaking rates is not required.

Soleymanpour et al. (2022) introduced a dysarthria embedding into FastSpeech 2 that allows to explicitly model and generate speech of different severity levels. They demonstrated that this synthetic dysarthric speech can augment the training data for dysarthric ASR on the Torgo corpus. We reproduce their findings on UA-Speech. We then ask whether dysarthric TTS could also be used to generate ASR training data for a new speaker based on just 5 or 100 recordings in a few-shot scenario.

6.3.1 Approach

In this section we describe the works on which our dysarthric TTS pipeline is based and any modifications that we have made.

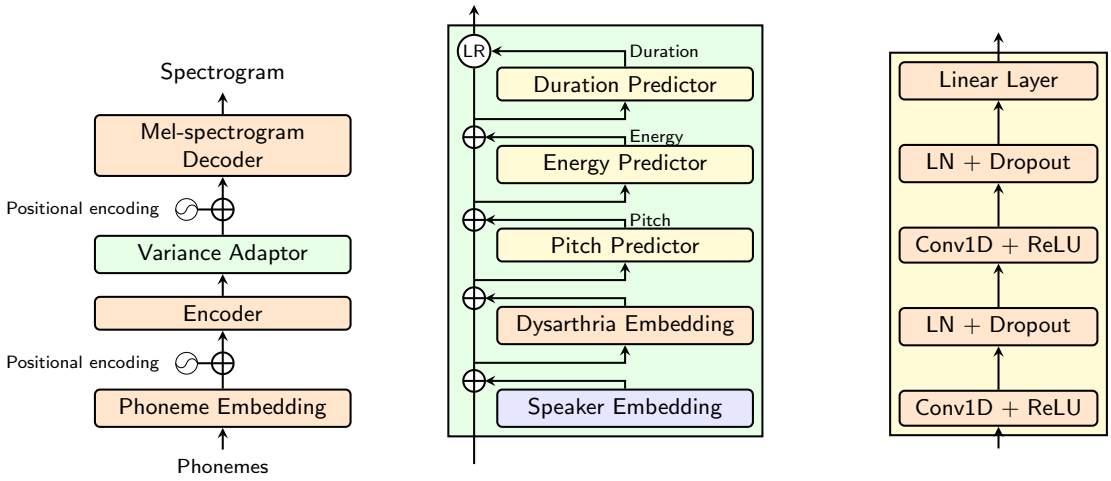
Controllable TTS

FastSpeech 2 (Ren et al., 2021) is a transformer-based non-autoregressive TTS system that allows for fast training and inference. Figure 6.6 illustrates the architecture of the model, which comprises a phoneme encoder and a Mel-spectrogram decoder. In between, it has a *variance adaptor* block to model different sources of variance in the speech signal and to control the TTS output. The variance adaptor contains multiple variance predictors. These are small neural networks that are trained to predict attributes like pitch, energy and phoneme duration. A length regulator expands the encoded input from phoneme- to frame-level based on the durations, while embeddings from the other predictors are added to the input. At training time, ground-truth values are used instead of the predictions. At inference time, the pitch, duration, and energy values can be adjusted to create variability in the synthesised speech.

The original FastSpeech 2 (Ren et al., 2021) predicts pitch spectrograms obtained from the continuous wavelet transform, but we use an implementation that directly predicts pitch values (Chien et al., 2021). We also follow their approach of placing the length regulator after all other variance predictors. Chien et al. (2021) also extended FastSpeech 2 to multiple speakers by adding a speaker embedding to the encoded input. The following variance predictors are thus conditioned on the speaker identity. We train this speaker embedding jointly with the rest of the network.

Dysarthric TTS

Soleymanpour et al. (2022) added a dysarthria severity predictor before the other variance predictors, so that their embeddings are conditioned on the severity of dysarthria of the speaker. Due to the controllable nature of FastSpeech 2, speech of different severity levels can then be generated, which they used for data augmentation in a dysarthric ASR system. As severity depends only on the speaker and cannot be predicted from text, we just use a severity embedding here that is learned jointly with the rest of the model, instead of a separate



(a) FastSpeech 2 (b) Variance adaptor (c) Variance predictors

Figure 6.6: Our adapted FastSpeech 2 architecture (figure adapted from Ren et al. (2021)). LR in subfigure (b) denotes the FastSpeech 2 length regulator and LN in subfigure (c) denotes layer normalisation. The speaker embeddings are obtained from a pretrained model and remain fixed.

predictor network. We group the speakers into the same 3 groups that each get their own embedding: unimpaired control speech, mild to moderate dysarthria, severe dysarthria.

Soleymanpour et al. (2022) developed heuristics to insert additional pauses into the synthetic speech for data augmentation experiments with continuous speech from the Torgo corpus. We do not add such heuristics because we only synthesise isolated words for UA-Speech.

Few-shot TTS

Chien et al. (2021) found that speaker embeddings from a generative VC system performed better than jointly trained ones or embeddings trained on a discriminative speaker classification task like x-vectors (Snyder et al., 2018). They chose embeddings from the AdaIN-VC system for one-shot voice conversion (Chou and Lee, 2019), so that the TTS would also support speakers not seen during training.

AdaIN-VC (Chou and Lee, 2019) is able to convert an utterance to an unseen speaker's voice from a single sample by separately encoding speaker and content. Speaker labels are not required for training, the speaker identity is assumed to be in the constant information throughout an utterance, while the content information is changing. An adaptive instance normalisation (AdaIN) (Huang and Belongie, 2017) layer means that no parameters have to be learned for a new speaker.

Following Soleymanpour et al. (2022), we train speaker embeddings jointly with the FastSpeech 2 model in the data augmentation experiments. This limits the set of speakers for which speech can be synthesised to those present in the training data. For the few-shot experiments, we have no such restriction because of the one-shot capable AdaIN-VC speaker embeddings and we investigate how little data is required from a target speaker to synthesise dysarthric speech and build a speaker-dependent ASR system for them.

6.3.2 Experimental setup

The audio files have a sampling rate of 16 kHz. For compatibility with existing code and pretrained models, we upsample the data to 22050 Hz in the TTS pipeline, while all ASR models are trained on 16 kHz.

TTS

We use synthetic speech for data augmentation, where we assume that training data for a target speaker is available, and in a few-shot setting, where we apply a trained TTS model on unseen speakers.

For data augmentation, we train one TTS model on all the training data from UA-Speech. For the few-shot experiments, we train 15 different models in a leave-one-speaker-out setup, i.e. on all control and the 14 other dysarthric speakers. We then use different amounts of dysarthric speech from blocks 1 and 3 of UA-Speech to obtain the speaker embeddings and as additional sources of ASR training data.

In each case, we train a phoneme-based FastSpeech 2 TTS model[4] with a batch size of 16 for 500k iterations in the default configuration. The input features are 80-dimensional Mel spectrograms. We obtain ground-truth phoneme durations for the duration predictor from forced alignment with a HMM/GMM ASR system trained on the same data. For the data augmentation experiments, speaker embeddings are trained jointly with the rest of the network. For the few-shot scenario, speaker embeddings are obtained from the AdaIN-VC model described in the next section.

For vocoding, we use the pretrained universal HiFi-GAN (Kong et al., 2020) model[5] and downsample its 22050 Hz output to 16 kHz for ASR training. We experimented with fine-tuning the vocoder on UA-Speech, but did not observe consistent benefits.

We again analyse the acoustic units of the synthesised dysarthric speech with the framework developed in Chapter 5.

[4] https://github.com/ming024/FastSpeech2
[5] https://github.com/jik876/hifi-gan

Speaker embeddings

For the few-shot scenario, we train AdaIN-VC models[6] on the same data as the TTS models with a batch size of 128 for 200k iterations using the default configuration, also with a leave-one-speaker-out setup. We train on the same Mel spectrograms that we extracted for FastSpeech 2 training, following Chou and Lee (2019). We take the 128-dimensional output of the speaker encoder as embeddings for FastSpeech 2 training and inference. We do not further fine-tune these embeddings during TTS training.

For the few-shot experiments, we select subsets of 5 and 100 words from the UA-Speech training blocks 1 and 3. We do not sample randomly, but instead choose words that offer the broadest phoneme coverage, emulating a scenario where target speakers are asked to record a small list of words with the biggest performance benefit. For each speaker, we pick a random utterance of each word, extract the AdaIN-VC embedding for it and take their average as the speaker embedding for speech synthesis, following Chou and Lee (2019). The TTS model is not trained or fine-tuned on these few-shot utterances, although fine-tuning could be explored in the future.

ASR

We train speaker-dependent LF-MMI acoustic models as described in Chapter 3 on only the data of the target dysarthric speaker, possibly augmented with synthetic speech. Although it is otherwise commonly done in LF-MMI training, we do not apply speed perturbation to the synthesised speech in these experiments because we can already manipulate the speed during TTS data augmentation.

6.3.3 Results and analysis

Data augmentation

As in the VC experiments, we first analyse the effect of the HiFiGAN vocoder by feeding the training data of both control and dysarthric speakers through the vocoder and training an LF-MMI model on it (*Both-HiFiGAN* in Table 6.3). Across dysarthric speakers, we only observe an absolute increase of 1.5% in WER. This is better than the MelGAN vocoder used for VC, indicating that this is the better option for future experiments.

For reference, we show the performance of an ASR system trained only on the control speech from UA-Speech (*Control*). The SD acoustic models trained on all speech of a given speaker from block 1 and 3 of UA-Speech represent the theoretical upper limit we can reach through data augmentation from a subset of that data. For comparison, we also show the results of SD models that additionally include all control speech (+ *Control*).

[6] https://github.com/cyhuang-tw/AdaIN-VC

Table 6.3: Word error rates (WER) on UA-Speech for each group of dysarthric speakers. For clarity, we also indicate whether the target speaker was seen during TTS training or not, where applicable (the baselines do not involve any TTS training).

Systems	Seen	Severe	Mod.-sev.	Moderate	Mild	Overall
Baselines from Chapter 3						
Dysarthric	–	62.4	32.2	29.2	19.0	34.0
Control	–	96.2	74.5	55.1	23.2	56.9
Both	–	62.8	35.7	28.7	17.7	33.9
Both-HiFiGAN	–	67.6	36.0	29.2	18.5	35.4
SD (Top-line)	–	70.3	42.7	38.2	24.0	41.3
+ Control	–	65.8	34.3	25.3	15.4	32.8
Data augmentation						
TTS-augmented	✓	70.8	38.7	33.6	18.5	37.6
+ Control	✓	67.7	32.6	26.1	14.9	32.8
TTS-augmented4x	✓	68.5	36.9	32.4	19.2	36.7
+ Control	✓	68.6	30.9	27.4	15.1	33.0
Few-shot						
F5-ctl	✗	99.6	99.1	98.1	92.0	96.5
+ Control	✗	94.9	75.9	55.8	22.3	56.7
F100-ctl	✗	98.8	99.0	92.5	83.3	91.7
+ Control	✗	93.8	75.6	51.7	21.8	55.4
F5-dys	✗	99.4	99.6	98.5	95.4	97.8
+ Control	✗	94.4	76.1	53.9	22.5	56.8
F100-dys	✗	99.3	99.2	95.6	91.3	95.6
+ Control	✗	94.5	72.7	52.4	20.6	54.6
F5-mix	✗	99.3	99.1	98.3	92.1	96.5
+ Control	✗	94.7	75.8	55.6	21.4	56.3
F100-mix	✗	98.6	97.1	92.7	82.6	91.3
+ Control	✗	93.7	72.7	50.7	20.9	54.2
Synthetic data only						
TTS-only	✓	98.2	93.9	87.8	85.1	90.5
TTS-only4x	✓	98.0	92.6	86.5	79.7	87.9

First, we generate a set of synthetic speech for data augmentation with the same words and number of utterances as the original UA-Speech training data. In order to generate multiple variants of the same word, we sample random pitch factors from {0.1, 0.6, 1.2, 1.75}, energy factors from {0.1, 1.0, 2.0}, duration factors from {1.0, 1.3, 1.6, 1.8} and the dysarthria embedding from {control, mild, severe} like Soleymanpour et al. (2022). First, we confirm their findings that augmenting the training data with synthetic dysarthric speech (*TTS-augmented*) improves speech recognition. We also confirm that adding four times as much synthetic speech further lowers the WER (*TTS-augmented4x*). However, when also adding the control speech itself, there is no further benefit from data augmentation.

Few-shot

We compare estimating the speaker embedding from 5 (*F5*) and 100 (*F100*) single-word utterances of the target speaker. These utterances are then also included for the training of the acoustic model. In either case, the total number of ASR training utterances is matched with the baseline. All of these models perform poorly with average WERs in the nineties, not even coming close to the control speech model. Nevertheless, we can observe certain patterns, e.g. estimating the speaker embedding from more utterances improves results.

We either set the dysarthria embedding to generate control speech (*F5*/*100-ctl*), speech of the same severity as the target speaker (*F5*/*100-dys*) or a mix of control, mild, and severely dysarthric speech (*F5*/*100-mix*). Curiously, we find that this mix or generating only control speech works better than matching the target severity. This could be because synthesising dysarthric speech introduces some dysarthria-like characteristics that are nonetheless not representative of the target speaker and more detrimental for ASR because the speaker embedding is only designed to capture general speaker information.

We also see slight improvements when combining the F100 data with control speech (+ *Control*). This indicates that while the synthetic speech on its own is not yet of sufficient quality, it can still yield benefits in combination with other data. To further evaluate this, we train another set of SD models on only the synthesised portion of the data used in the *TTS-augmented* experiments, where the target speakers were already seen during TTS training (*TTS-only*). Indeed, even these results are very poor although the speakers were seen and the TTS output was beneficial for ASR data augmentation. This suggests that no significant improvements can be expected in the few-shot setting before the TTS quality in general is not further increased. Listening to synthesised speech samples also indicates that the TTS model sometimes has difficulties finding the right alignment between phonemes and the audio.

Analysis

We evaluate the quality of the synthetic dysarthric speech by analysing the discriminability of the monophone acoustic units. In the future, it would be also worthwhile to apply the objective evaluation measures proposed by Halpern et al. (2021). Figure 6.7 (a) shows the relationship between median KL divergences of the synthetic speech used in the data augmentation experiments for each dysarthric speaker and their subjective intelligibility ratings (Pearson's r = 0.85), compared with the original dysarthric speech (r = 0.90). In terms of acoustic space discriminability, the synthetic speech is correctly showing the same patterns as the original dysarthric speech.

For data augmentation, we synthesised speech with the dysarthria embedding set to a different random value for each utterance. But how does the TTS output change when we set the dysarthria embedding to generate control, mild, or severely dysarthric speech? For each embedding value, we synthesise one utterance for each word in the UA-Speech training data.

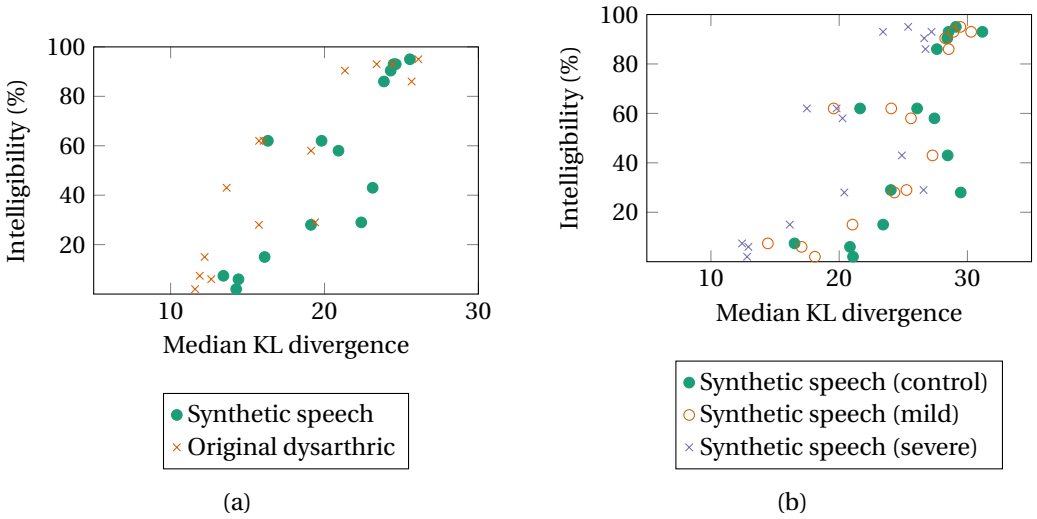


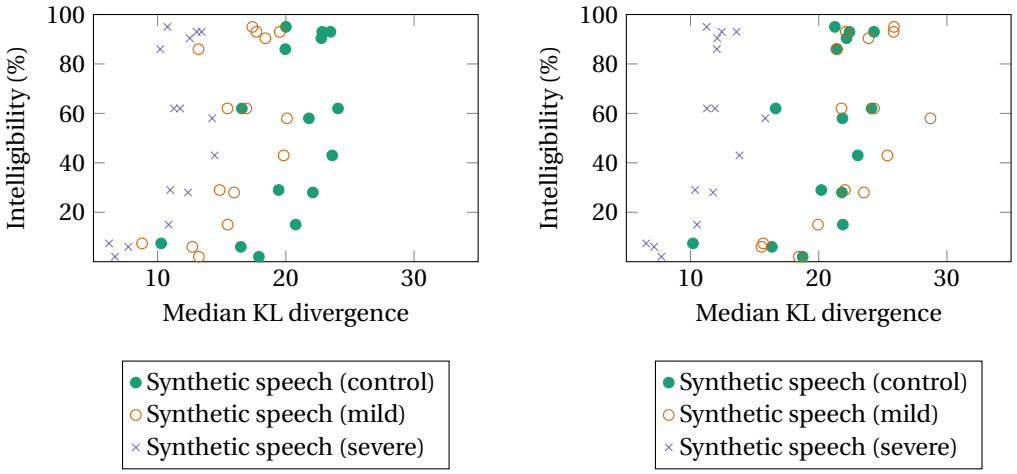
Figure 6.7: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings of (a) synthetic dysarthric speech used for data augmentation experiments on UA-Speech (Pearson's $r = 0.85$) and of the original dysarthric speech ($r = 0.90$). (b) shows the same relationship for speech synthesised with the dysarthria embedding set to control, mild, or severe.

We find that the dysarthria embedding learns to correctly influence the length regulator, with average utterance durations of 1.2s for control, 1.9s for mildly dysarthric and 2.6s for severely dysarthric synthesised speech. Figure 6.7 (b) shows the relationship between median KL divergences of these three sets of synthesised speech and the subjective intelligibility ratings of each dysarthric speaker. Indeed, the median KL divergences decrease for mild and severely dysarthric synthesised speech, indicating reduced discriminability. We note that when synthesising with the dysarthria embedding set to *control*, there is still a correlation between median KL divergences and subjective intelligibility ratings. This is due to the speaker embedding that inevitably also captures dysarthria characteristics of the speaker, so it is not expected that this synthesised control speech sounds like a control speaker without dysarthria.

However, in the few-shot experiments we synthesised speech for new speakers that were not seen during TTS training. We again generate a set of control, mild, and severely dysarthric speech by setting the dysarthria embedding accordingly with the few-shot model for each unseen speaker. Figure 6.8 shows that there are meaningful differences in the acoustic space between the three severity levels for these unseen speakers as well, both for speaker embeddings obtained from 5 and from 100 utterances of that speaker.

6.4 Summary

The aim of this chapter was to compare different data augmentation methods both in terms of ASR performance and the characteristics of the generated speech. In all cases, VC with



(a) 5-shot

(b) 100-shot

Figure 6.8: Relationship between the median KL divergence of acoustic units and subjective intelligibility ratings of synthetic dysarthric speech used for few-shot experiments on UA-Speech with the dysarthria embedding set to control, mild, or severe. The speaker embeddings are estimated from (a) 5 and (b) 100 utterances from the target speaker, but their speech is not included during TTS training.

signal processing and GAN-based models and TTS, our results suggest that the overall acoustic quality of the generated outputs is not sufficient for accurate speech recognition yet. While it helped to add the generated data to dysarthric speech, it was even better to just add the original control speech. For GAN-based VC, the poor results could to some degree be explained by the effect of the vocoding step.

However, we found that both GAN-based VC and TTS learn to model dysarthric speech characteristics and reproduce differences in acoustic space discriminability between speakers of different severity that are observed in the original dysarthric speech.

7 Conclusions

7.1 Conclusions

In this thesis we investigated methods to improve dysarthric ASR systems and make them more resistant to the acoustic and lexical variability of dysarthric speech by considering the relationship between the acoustic model and its training data.

We developed strong baselines for the Torgo and UA-Speech corpora of dysarthric speech based on sequence-discriminative LF-MMI training. We found that frame subsampling and data augmentation with speed perturbation are important factors for their success. The resulting models in particular reduce insertion errors, which can otherwise be frequent due to the low speaking rates of dysarthric speakers.

We observed that adding unimpaired control speech to the training data is always beneficial for dysarthric ASR. However, for LF-MMI training, it led to worse results on control speakers compared to models trained only on control speech or to SGMMs. We were able to compensate for this by dynamically combining acoustic models trained on different groups of speakers. This model combination approach also improved WERs of dysarthric speakers. We further extended it to the combination of models trained with either phoneme or grapheme acoustic units in order to implicitly handle pronunciation variants in dysarthric speech.

To better understand the differences in performance of different acoustic models, we proposed an analysis framework based on the acoustic discriminability of the training data by computing KL divergences between Gaussian distributions estimated for each acoustic unit. When comparing dysarthric speakers, this analysis showed high correlations between discriminability of the acoustic unit space and both subjective intelligibility ratings and ASR results of SD acoustic models, underlining the viability of this approach.

Finally, we compared multiple data augmentation approaches for dysarthric ASR, including VC and TTS, within this analysis framework. We observed that synthetic dysarthric speech at different severity levels generated with both GAN-based VC and TTS shows similar effects

on the acoustic unit space as the original dysarthric speech. Data augmentation with the generated speech also improved WERs with respect to acoustic models trained only on the original dysarthric speech. However, in each case we found it to be even better to just augment with the original control speech instead, concluding that the output quality of these systems needs to be further improved before they are applied to dysarthric ASR.

7.2 Directions for future research

We suggest the following directions for future research:

- In this thesis we restricted ourselves to sequence-discriminative LF-MMI training in the hybrid HMM/DNN ASR framework. In the meantime, end-to-end approaches, such as CTC or RNN transducer (RNN-T), have gained in popularity and competitiveness. They could be compared to the acoustic models presented in this work. They also facilitate the integration of other types of acoustic features, such as pretrained speech representations, and multi-task training.
- Our work has mainly focused on dysarthric ASR, but it could be extended to other speech pathologies. While we also briefly evaluated our model combination approach to handle lexical variability in children's and non-native children's speech, more comparisons with other kinds of atypical speech could be carried out. Similarly, our analysis framework should be validated on other forms of atypical speech as well. For example, it could analyse acoustic differences in children's, elderly, or accented speech.
- VC and TTS data augmentation for dysarthric ASR proved to be promising in that the generated speech samples successfully reproduced characteristics of the original dysarthric speech. It turned out to be better to just augment the training data with the original control speech than any voice-converted or synthesised samples, but further improvements in the output quality of these systems are possible. The vocoding also had a negative effect on ASR performance and in the future this step could be circumvented by training acoustic models directly on the generated Mel spectrograms.
- Few-shot dysarthric ASR approaches are likely to become more widespread. We concluded that adding control speech when training acoustic models on dysarthric speech is always beneficial. But we can consider the opposite question: how little dysarthric speech is required to adapt a general purpose ASR system trained on larger corpora to an unseen speaker with dysarthria? This could gain wider adoption than training dysarthria-specific models from scratch because typical speech models and data are widely available.

Bibliography

Aradilla, G., Bourlard, H., and Magimai.-Doss, M. (2008). Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task. In *Proceedings of Interspeech*, pages 928–931.

Aradilla, G., Vepa, J., and Bourlard, H. (2007). An acoustic model based on Kullback-Leibler divergence for posterior features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 657–660.

Bahl, L., Brown, P., de Souza, P., and Mercer, R. L. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of ICASSP*, pages 49–52.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF_STAR Children's Speech Corpus. In *Proceedings of Interspeech*, pages 2761–2764.

Bell, P., Gales, M. J. F., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., and Woodland, P. C. (2015). The MGB challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 687–693.

Benoît, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381–392.

Bhat, C., Das, B., Vachhani, B., and Kopparapu, S. K. (2018). Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In *Proceedings of Interspeech*, pages 451–455.

Boersma, P. and Weenink, D. (2021). Praat: doing phonetics by computer (version 6.1.38). Retrieved from http://www.praat.org/.

Bourlard, H. A. and Morgan, N. (1994). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y. (2021). Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech. In *Proceedings of ICASSP*, pages 8588–8592.

Chou, J.-c. and Lee, H.-Y. (2019). One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *Proceedings of Interspeech*, pages 664–668.

Chow, Y.-L. (1990). Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm. In *Proceedings of ICASSP*, pages 701–704.

Christensen, H., Aniol, M. B., Bell, P., Green, P., Hain, T., King, S., and Swietojanski, P. (2013). Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *Proceedings of Interspeech*, pages 3642–3645.

Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *Proceedings of Interspeech*, pages 1776–1779.

Dodd, B., Holm, A., Hua, Z., and Crosbie, S. (2003). Phonological development: a normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, 17(8):617–643.

Dubagunta, S. P., Kabil, S. H., and Magimai.-Doss, M. (2019). Improving Children Speech Recognition through Feature Learning from Raw Speech Signal. In *Proceedings of ICASSP*, pages 5736–5740.

Dubagunta, S. P., van Son, R., and Doss, M. M. (2020). Adjustable Deterministic Pseudonymisation of Speech: Idiap-NKI's submission to VoicePrivacy 2020 Challenge. In *Proceedings of the VoicePrivacy Challenge*.

Dubagunta, S. P., van Son, R. J., and Magimai.-Doss, M. (2022). Adjustable deterministic pseudonymization of speech. *Computer Speech & Language*, 72.

Duffy, J. R. (2012). *Motor Speech Disorders*. Mosby, third edition.

Durrieu, J.-L., Thiran, J.-P., and Kelly, F. (2012). Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian Mixture Models. In *Proceedings of ICASSP*, pages 4833–4836.

Enderby, P. (1980). Frenchay Dysarthria Assessment. *International Journal of Language & Communication Disorders*, 15(3):165–173.

España-Bonet, C. and Fonollosa, J. A. R. (2016). Automatic Speech Recognition with Deep Neural Networks for Impaired Speech. In *Proceedings of IberSpeech*, pages 97–107.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of ASRU*, pages 347–354.

Fosler-Lussier, E. (1999a). Contextual Word and Syllable Pronunciation Models. In *Proceedings of ASRU*.

Fosler-Lussier, E. (1999b). Multi-level Decision Trees for Static and Dynamic Pronunciation Models. In *Proceedings of Eurospeech*.

Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

Gales, M., Knill, K., and Ragni, A. (2015). Unicode-based graphemic systems for limited resource languages. In *Proceedings of ICASSP*, pages 5186–5190.

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., and Meng, H. (2020). Investigation of Data Augmentation Techniques for Disordered Speech Recognition. In *Proceedings of Interspeech*, pages 696–700.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of ICML*, pages 369–376.

Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., Nelson, P. C., and Tomanek, K. (2021). Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In *Proceedings of Interspeech*, pages 4778–4782.

Gretter, R., Matassoni, M., Bannò, S., and Daniele, F. (2020). TLT-school: A Corpus of Non Native Children Speech. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 378–385.

Gretter, R., Matassoni, M., Falavigna, D., Misra, A., Leong, C., Knill, K., and Wang, L. (2021). ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. In *Proceedings of Interspeech*, pages 3845–3849.

Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018a). End-to-end speech recognition using lattice-free MMI. In *Proceedings of Interspeech*, pages 12–16.

Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018b). Flat-start Single-stage Discriminatively Trained HMM-based Models for ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):1949–1961.

Hahm, S., Heitzman, D., and Wang, J. (2015). Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-Speaker Articulatory Normalization. In *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 47–54.

Halpern, B. M., Fritsch, J., Hermann, E., van Son, R., Scharenborg, O., and Magimai.-Doss, M. (2021). An Objective Evaluation Framework for Pathological Speech Synthesis. In *Proceedings of the ITG Conference on Speech Communication*.

Harvill, J., Issa, D., Hasegawa-Johnson, M., and Yoo, C. (2021). Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary. In *Proceedings of ICASSP*, pages 6428–6432.

Hermann, E. and Magimai.-Doss, M. (2020). Dysarthric Speech Recognition with Lattice-Free MMI. In *Proceedings of ICASSP*, pages 6109–6113.

Hermann, E. and Magimai.-Doss, M. (2021). Handling acoustic variation in dysarthric speech recognition systems through model combination. In *Proceedings of Interspeech*, pages 4788–4792.

Hermann, E. and Magimai.-Doss, M. (2023). Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation. In *Proceedings of Interspeech (accepted)*.

Hermansky, H. (2013). Multistream Recognition of Speech: Dealing With Unknown Unknowns. *Proceedings of the IEEE*, 101(5):1076–1088.

Hernandez, A., Pérez-Toro, P. A., Noeth, E., Orozco-Arroyave, J. R., Maier, A., and Yang, S. H. (2022). Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. In *Proceedings of Interspeech*, pages 51–55.

Hinsvark, A., Delworth, N., Del Rio, M., McNamara, Q., Dong, J., Westerman, R., Huang, M., Palakapilly, J., Drexler, J., Pirkin, I., Bhandari, N., and Jette, M. (2021). Accented Speech Recognition: A Survey. Technical Report arXiv:2104.10747.

Hu, S., Liu, S., Xie, X., Geng, M., Wang, T., Hu, S., Cui, M., Liu, X., and Meng, H. (2022). Exploiting Cross Domain Acoustic-to-Articulatory Inverted Features for Disordered Speech Recognition. In *Proceedings of ICASSP*, pages 6747–6751.

Huang, W.-C., Halpern, B. M., Phillip Violeta, L., Scharenborg, O., and Toda, T. (2022). Towards Identity Preserving Normal to Dysarthric Voice Conversion. In *Proceedings of ICASSP*, pages 6672–6676. IEEE.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1501–1510.

Illa, M., Halpern, B. M., van Son, R., Moro-Velazquez, L., and Scharenborg, O. (2021). Pathological voice adaptation with autoencoder-based voice conversion. In *Proceedings of the ISCA Speech Synthesis Workshop (SSW)*, pages 19–24.

Irie, K., Prabhavalkar, R., Kannan, A., Bruguier, A., Rybach, D., and Nguyen, P. (2019). On the Choice of Modeling Unit for Sequence-to-Sequence Speech Recognition. In *Proceedings of Interspeech*, pages 3800–3804.

Jiao, Y., Tu, M., Berisha, V., and Liss, J. (2018). Simulating Dysarthric Speech for Training Data Augmentation in Clinical Speech Applications. In *Proceedings of ICASSP*, pages 6009–6013.

Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., and Liu, X. (2022a). Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition. Technical Report arXiv:2205.06445.

Jin, Z., Geng, M., Xie, X., Yu, J., Liu, S., Liu, X., and Meng, H. (2021). Adversarial Data Augmentation for Disordered Speech Recognition. In *Proceedings of Interspeech*, pages 4803–4807.

Jin, Z., Xie, X., Geng, M., Wang, T., Hu, S., Deng, J., Li, G., and Liu, X. (2022b). Adversarial Data Augmentation Using VAE-GAN for Disordered Speech Recognition. Technical Report arXiv:2211.01646.

Joy, N. M. and Umesh, S. (2018). Improving acoustic models in TORGO dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):637–645.

Kadyan, V., Kathania, H., Govil, P., and Kurimo, M. (2021). Synthesis Speech Based Data Augmentation for Low Resource Children ASR. In Karpov, A. and Potapova, R., editors, *Speech and Computer*, volume 12997, pages 317–326. Springer, Cham.

Kamper, H. (2019). Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In *Proceedings of ICASSP*, pages 6535–6539.

Kaneko, T. and Kameoka, H. (2018). CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019a). Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion. In *Proceedings of ICASSP*, pages 6820–6824. IEEE.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019b). StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion. In *Proceedings of Interspeech*, pages 679–683.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2021). MaskCycleGAN-VC: Learning Non-Parallel Voice Conversion with Filling in Frames. In *Proceedings of ICASSP*, pages 5919–5923. IEEE.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric Speech Database for Universal Access Research. In *Proceedings of Interspeech*, pages 1741–1744.

Kim, M., Wang, J., and Kim, H. (2016). Dysarthric Speech Recognition Using Kullback-Leibler Divergence-Based Hidden Markov Model. In *Proceedings of Interspeech*, pages 2671–2675.

King, S. and Karaiskos, V. (2016). The Blizzard Challenge 2016. In *Proceedings of the Blizzard Challenge Workshop*.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. In *Proceedings of Interspeech*, pages 3586–3589.

Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proceedings of NeurIPS*.

Kumar, K. and Kumar, R. (2019). MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Proceedings of NeurIPS*.

Lansford, K. L. and Liss, J. M. (2014). Vowel Acoustics in Dysarthria: Speech Disorder Diagnosis and Classification. *Journal of Speech, Language, and Hearing Research*, 57(1):57–67.

Lee, C.-H. and Gauvain, J.-L. (1993). Speaker adaptation based on MAP estimation of HMM parameters. In *Proceedings of ICASSP*, pages 558–561 vol.2.

Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Liu, S., Geng, M., Hu, S., Xie, X., Cui, M., Yu, J., Liu, X., and Meng, H. (2021). Recent Progress in the CUHK Dysarthric Speech Recognition System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2267–2281.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Mengistu, K. T. and Rudzicz, F. (2011a). Adapting Acoustic and Lexical Models to Dysarthric Speech. In *Proceedings of ICASSP*, pages 4924–4927.

Mengistu, K. T. and Rudzicz, F. (2011b). Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech. In *Proceedings of Canadian Conference on Artificial Intelligence*, pages 291–300.

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted Finite-State Transducers in Speech Recognition. *Computer Speech and Language*, 16(1):69–88.

Moore, M., Saxon, M., Venkateswara, H., Berisha, V., and Panchanathan, S. (2019). Say What? A Dataset for Exploring the Error Patterns That Two ASR Engines Make. In *Proceedings of Interspeech*, pages 2528–2532.

Moore, M., Venkateswara, H., and Panchanathan, S. (2018). Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems. In *Proceedings of Interspeech*, pages 466–470.

Mustafa, M. B., Salim, S. S., Al-Qatab, M. N., and Siong, B. E. (2014). Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers. *PLoS ONE*, 9(1):1–11.

Nicolao, M., Christensen, H., Cunningham, S., Green, P., and Hain, T. (2016). A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus. In *Proceedings of LREC*, pages 1993–1997.

Novak, J. R., Minematsu, N., and Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

Palaz, D., Magimai.-Doss, M., and Collobert, R. (2019). End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition. *Speech Communication*, 108:15–32.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Park, D. S., Zhang, Y., Chiu, C.-C., Chen, Y., Li, B., Chan, W., Le, Q. V., and Wu, Y. (2020). Specaugment on Large Scale Datasets. In *Proceedings of ICASSP*, pages 6879–6883.

Povey, D. (2003). *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R. C., Schwarz, P., and Thomas, S. (2010). The subspace Gaussian mixture model - A structured model for speech recognition. *Computer Speech & Language*, 25:404–439.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proceedings of Interspeech*, pages 3743–3747.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of Interspeech*, pages 2751–2755.

Prananta, L., Halpern, B. M., Feng, S., and Scharenborg, O. (2022). The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition. In *Proceedings of Interspeech*, pages 36–40.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rasipuram, R., Bell, P., and Magimai.-Doss, M. (2013). Grapheme and multilingual posterior features for under-resourced speech recognition: A study on Scottish Gaelic. In *Proceedings of ICASSP*, pages 7334–7338.

Rasipuram, R. and Doss, M. M. (2012). Combining Acoustic Data Driven G2P and Letter-to-Sound Rules for Under Resource Lexicon Generation. In *Proceedings of Interspeech*, pages 1820–1823.

Rasipuram, R. and Magimai.-Doss, M. (2015). Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model. *Speech Communication*, 68:23–40.

Razavi, M. and Magimai.-Doss, M. (2015). An HMM-based formalism for automatic subword unit derivation and pronunciation generation. In *Proceedings of ICASSP*, pages 4639–4643.

Razavi, M., Rasipuram, R., and Magimai.-Doss, M. (2018). Towards weakly supervised acoustic subword unit discovery and lexicon development using hidden Markov models. *Speech Communication*, 96:168–183.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition. In *Proceedings of ICASSP*, pages 81–84.

Rudzicz, F. (2011). Articulatory Knowledge in the Recognition of Dysarthric Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960.

Rudzicz, F., Hirst, G., and Lieshout, P. V. (2012a). Vocal Tract Representation in the Recognition of Cerebral Palsied Speech. *Journal of Speech, Language and Hearing Research*, 55(August):1190–1207.

Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012b). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources & Evaluation*, 46(4):523–541.

Sak, H., Senior, A., Rao, K., and Beaufays, F. (2015). Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. In *Proceedings of Interspeech*, pages 1468–1472.

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker Adaptation of Neural Network Acoustic Models Using I - Vectors. In *Proceedings of ASRU*, pages 55–59.

Saraçlar, M. and Khudanpur, S. (2004). Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech & Language*, 18(4):375–395.

Schu, G., Janbakhshi, P., and Kodrasi, I. (2023). On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. In *Proceedings of ICASSP (to appear)*.

Sharma, H. V. and Hasegawa-Johnson, M. (2010). State-Transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition. In *Proceedings of SLPAT*, pages 72–79.

Sisman, B., Yamagishi, J., King, S., and Li, H. (2021). An Overview of Voice Conversion and its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of ICASSP*, pages 5329–5333.

Soleymanpour, M., Johnson, M. T., Soleymanpour, R., and Berry, J. (2022). Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 7382–7386.

Sriranjani, R., Umesh, S., and Ramasubba Reddy, M. (2015). Pronunciation Adaptation For Disordered Speech Recognition Using State-Specific Vectors of Phone-Cluster Adaptive Training. In *Proceedings of SLPAT*, pages 72–78.

Swietojanski, P. and Renals, S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, pages 171–176. IEEE.

Tejedor, J., Wang, D., Frankel, J., King, S., and Colás, J. (2008). A comparison of grapheme and phoneme-based units for Spanish spoken term detection. *Speech Communication*, 50(11-12):980–991.

Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking: Speech chain by deep learning. In *Proceedings of ASRU*, pages 301–308.

Tjandra, A., Sakti, S., and Nakamura, S. (2020). Machine Speech Chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.

Tobin, J. and Tomanek, K. (2022). Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets. In *Proceedings of ICASSP*, pages 6637–6641.

Tomanek, K., Zayats, V., Padfield, D., Vaillancourt, K., and Biadsy, F. (2021). Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6751–6760.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., and Maouche, M. (2022). The VoicePrivacy 2020 Challenge: Results and findings. *Computer Speech & Language*, 74.

Tripoliti, E., Zrinzo, L., Martinez-Torres, I., Frost, E., Pinto, S., Foltynie, T., Holl, E., Petersen, E., Roughton, M., Hariz, M. I., and Limousin, P. (2011). Effects of subthalamic stimulation on speech of consecutive patients with Parkinson disease. *Neurology*, 76(1):80–86.

Turner, G. S., Tjaden, K., and Weismer, G. (1995). The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals With Amyotrophic Lateral Sclerosis. *Journal of Speech, Language, and Hearing Research*, 38(5):1001–1013.

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., and Badino, L. (2021). EasyCall corpus: A dysarthric speech dataset. In *Proceedings of Interspeech*, pages 41–45.

Tykalová, T., Rusz, J., Čmejla, R., Klempíř, J., Růžičková, H., Roth, J., and Růžička, E. (2015). Effect of dopaminergic medication on speech dysfluency in Parkinson's disease: A longitudinal study. *Journal of Neural Transmission*, 122(8):1135–1142.

Vachhani, B., Bhat, C., and Kopparapu, S. K. (2018). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In *Proceedings of Interspeech*, pages 471–475.

Valtchev, V., Odell, J., Woodland, P., and Young, S. (1996). Lattice-based discriminative training for large vocabulary speech recognition. In *Proceedings of ICASSP*, volume 2, pages 605–608.

Wang, D., Yu, J., Wu, X., Sun, L., Liu, X., and Meng, H. (2021). Improved End-to-End Dysarthric Speech Recognition via Meta-learning Based Model Re-initialization. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Wang, T., Deng, J., Geng, M., Ye, Z., Hu, S., Wang, Y., Cui, M., Jin, Z., Liu, X., and Meng, H. (2022). Conformer Based Elderly Speech Recognition System for Alzheimer's Disease Detection. In *Proceedings of Interspeech*, pages 4825–4829.

Xiao, A., Zheng, W., Keren, G., Le, D., Zhang, F., Fuegen, C., Kalinli, O., Saraf, Y., and Mohamed, A. (2022). Scaling ASR Improves Zero and Few Shot Learning. In *Proceedings of Interspeech*, pages 5135–5139.

Xiong, F., Barker, J., and Christensen, H. (2018). Deep Learning of Articulatory-Based Representations and Applications for Improving Dysarthric Speech Recognition. In *Proceedings of the ITG Conference on Speech Communication*, pages 331–335.

Xiong, F., Barker, J., and Christensen, H. (2019). Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 5836–5840.

Xiong, F., Barker, J., Yue, Z., and Christensen, H. (2020). Source Domain Data Selection for Improved Transfer Learning Targeting Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 7424–7428.

Xu, H., Povey, D., Mangu, L., and Zhu, J. (2011). Minimum Bayes Risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, 25(4):802–828.

Yılmaz, E., Ganzeboom, M., Cucchiarini, C., and Strik, H. (2016). Combining Non-pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech. In *Proceedings of Interspeech*, pages 218–222.

Yılmaz, E., Mitra, V., Bartels, C., and Franco, H. (2018). Articulatory Features for ASR of Pathological Speech. In *Proceedings of Interspeech*, pages 2958–2962.

Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the Workshop on Human Language Technology (HLT)*, pages 307–312.

Yu, J., Xie, X., Liu, S., Hu, S., Lam, M. W. Y., Wu, X., Wong, K. H., Liu, X., and Meng, H. (2018). Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus. In *Proceedings of Interspeech*, pages 2938–2942.

Yue, Z., Loweimi, E., Christensen, H., Barker, J., and Cvetkovic, Z. (2022a). Acoustic Modelling From Raw Source and Filter Components for Dysarthric Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2968–2980.

Yue, Z., Loweimi, E., and Cvetkovic, Z. (2022b). Raw Source and Filter Modelling for Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 7377–7381.

Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., and Barker, J. (2022c). Multi-Modal Acoustic-Articulatory Feature Fusion For Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 7372–7376.

Yue, Z., Xiong, F., Christensen, H., and Barker, J. (2020). Exploring Appropriate Acoustic and Language Modelling Choices for Continuous Dysarthric Speech Recognition. In *Proceedings of ICASSP*, pages 6094–6098.

Zhang, C. and Woodland, P. C. (2016). DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions. In *Proceedings of ICASSP*, pages 5300–5304. IEEE.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of ICCV*, pages 2242–2251. IEEE.

Glossary

ALS amyotrophic lateral sclerosis

ASR automatic speech recognition

CE cross-entropy

CNN convolutional neural network

CTC connectionist temporal classification

DNN deep neural network

DTW dynamic time warping

GAN generative adversarial network

GMM Gaussian mixture model

HMM hidden Markov model

KL Kullback-Leibler

KL-HMM Kullback-Leibler divergence based hidden Markov model

LF-MMI lattice-free maximum mutual information

LHUC learning hidden unit contributions

LM language model

MAP maximum a posteriori

MBR minimum Bayes risk

MFCC Mel-frequency cepstral coefficient

MLLR maximum likelihood linear regression

RNN recurrent neural network

RNN-T RNN transducer

SAT speaker adaptive training

SD speaker-dependent

SGMM subspace GMM

SI speaker-independent

TDNN time-delay neural network

TTS text-to-speech

VC voice conversion

WER word error rate

Enno Hermann

enno.hermann@idiap.ch enno.xyz eginhard ennoh

Research Interests

Automatic speech recognition, speech synthesis, deep learning, low-resource speech processing.

Education

2018 – 2023 **École polytechnique fédérale de Lausanne (EPFL), Switzerland**
PhD, Electrical Engineering
Supervisors: Dr. Jean-Marc Odobez, Dr. Mathew Magimai.-Doss
Thesis: *On matching data and model in LF-MMI-based dysarthric speech recognition*

2016 – 2017 **University of Edinburgh, United Kingdom**
MSc (*Distinction*), Speech and Language Processing
Supervisor: Prof. Sharon Goldwater
Thesis: *Iteratively improving unsupervised term discovery and unsupervised speech representations*

2012 – 2016 **Trinity College Dublin, Ireland**
BA (*First Class Honours*), Computer Science, Linguistics and French
Supervisor: Dr. Christer Gobl
Thesis: *Exploring bilingual text-to-speech conversion for Irish and Irish English using a statistical parametric synthesis system*

Experience

2023 – present **Idiap Research Institute, Switzerland** – *Postdoctoral Researcher*
Multilingual and emotional speech synthesis, speech recognition.

Summer 2021 **Therapy Box, United Kingdom*** – *Research Intern*
Syllable counting, speaker diarisation, phoneme recognition.

2018 – 2023 **Idiap Research Institute, Switzerland** – *Research Assistant*
Pathological speech recognition.

2017 – 2018 **University of Edinburgh, United Kingdom** – *Research Assistant*
Multilingual approaches to zero-resource speech processing.

Summer 2016 **Google, USA** – *Software Engineer Intern*
Unsupervised coreference resolution.

Summer 2015 **Google, United Kingdom** – *Software Engineer Intern*
Wrote a database and querying library for analysing voice corpora.

Summer 2014 **Google, Ireland** – *Site Reliability Engineer Intern*
Worked on a monitoring service.

Publications

- **Enno Hermann** and Mathew Magimai.-Doss. 'Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation'. In: *Proceedings of Interspeech (accepted)*. 2023.
- Timothy Piton, **Enno Hermann**, Angela Pasqualotto, Marjolaine Cohen, Mathew Magimai.-Doss and Daphné Bavelier. 'Using Commercial ASR Solutions to Assess Reading Skills in Children: A Case Report'. In: *Proceedings of Interspeech (accepted)*. 2023.

*remotely from Switzerland

- Bence Mark Halpern, Julian Fritsch, **Enno Hermann**, Rob van Son, Odette Scharenborg and Mathew Magimai.-Doss. 'An Objective Evaluation Framework for Pathological Speech Synthesis'. In: *Proceedings of the ITG Conference on Speech Communication*. 2021.
- **Enno Hermann**, Herman Kamper and Sharon Goldwater. 'Multilingual and Unsupervised Subword Modeling for Zero-Resource Languages'. In: *Computer Speech & Language* 65 (2021).
- **Enno Hermann** and Mathew Magimai.-Doss. 'Dysarthric Speech Recognition with Lattice-Free MMI'. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2020, pp. 6109–6113.
- **Enno Hermann** and Mathew Magimai.-Doss. 'Handling Acoustic Variation in Dysarthric Speech Recognition Systems Through Model Combination'. In: *Proceedings of Interspeech*. 2020, pp. 4788–4792.
- **Enno Hermann** and Sharon Goldwater. 'Multilingual Bottleneck Features for Subword Modeling in Zero-Resource Languages'. In: *Proceedings of Interspeech*. 2018, pp. 2668–2672.

Skills

Languages	German, English, French
Programming	Python, Bash, C++
Frameworks	Kaldi, PyTorch, LaTeX, Emacs, Git

Academic Service

Reviewing	ICASSP (2021, 2023), Interspeech (2021-2023), SLPAT (2022).
Teaching	Teaching assistant for: • Automatic Speech Processing, EPFL (2022-2023) • Introduction to Speech Processing, UniDistance (2019-2023) • Cognitive Science, University of Edinburgh (2018) • Accelerated Natural Language Processing, University of Edinburgh (2017)

Awards

2017 **Highly Commended Dissertation Prize**, School of Philosophy, Psychology & Language Sciences, University of Edinburgh.

2014 **Foundation Scholarship**, Trinity College Dublin.